Apnoea Detection

A thesis presented for the degree of
Doctor of Philosophy
in Electrical and Electronic Engineering
from the
University of Canterbury
Christchurch, New Zealand

by

Paul Michael Macey
B.E. (Hons)

April 1998

Abstract

This thesis is concerned with the detection of apnoeas in infants from an abdominal breathing signal, where an apnoea is a pause in breathing during sleep. Apnoea detection is performed by analysing breathing signals recorded during sleep studies. An abdominal breathing signal recorded by the BabyLog polysomnographic system is used for this research. A reference set of apnoeas is formed by three human experts identifying apnoeas five seconds and longer within ten overnight recordings of breathing. There was a 10% disagreement on the identification of events. Based on this reference set, the performances of existing methods of apnoea detection were evaluated, and found to have low incidence of false negatives but high incidence of false positives. An existing algorithm was developed, and an application of this algorithm as part of a study of low risk infants is presented. Properties that represent most apnoeas as found in an abdominal breathing signal are described. Human experts are consulted to determine what properties of the signal they use to recognise apnoeas, and a collection of deterministic, or shape, properties is condensed to represent expert opinion. An apnoea is modeled as a flat region with four properties: flatness, duration, thinness and smoothness. Mathematical descriptions that discriminate between apnoea and non-apnoea events of each property are formulated. An expert system for the classification of events is then developed, based on property measures being classified by a neural network. The system has achieved 95% to 98% accuracy for a false detection rate of 15% to 40%. Applications include scoring apnoeas for sleep studies, an aid to clinicians in diagnosing breathing problems, and developing standard definitions of breathing signals corresponding to apnoeas.

Acknowledgments

I completed this thesis with the support and encouragement of many people.

I am grateful to Dr. Rodney Ford who has continually supported me, as well as providing the research environment, financial backing, and direction in terms of research. I am also grateful to Dr. Jimmy Li, my supervisor for more than three years. Dr. Jimmy Li has trained me to do engineering research, and greatly contributed to the quality of this thesis in terms of the content and the presentation.

I also thank my other supervisors. My first supervisor Dr. Kathy Garden encouraged me into postgraduate studies and supported me during the early stages of my research, giving me guidance in terms of topics for my thesis. Prof. Jim Reilly supervised me for six months and he helped me understand what was required for a Ph.D. thesis. Dr. Phil Bones kindly agreed to supervise me in my final months, and has smoothed the administrative process greatly.

I worked with Dr. David Tappin, the Canterbury Cot Death Fellow, for over a year and greatly enjoyed his openness to new ideas and commitment to producing quality research. Craig Tuffnell and I have worked alongside for many years and developed a friendship which we enjoy to this day. I am very grateful to Craig for his help throughout, from when I first arrived in the BabyLog team, through to the final stages during which he has read my thesis thoroughly. Philip Schluter has helped with statistical analyses and also read some chapters, adding rigour to the results and writing. I have enjoyed working with the other members of the BabyLog team—Richard Dove, Janine Larkin, Brent Price, Richard Fright and Jeff Brown—fixing computers, programming and spending hours staring at computer screens! I have received guidance from the team in the areas of the data collection, computing hardware, software, and medical interpretation, to name a few. At the Community Paediatric Unit, where I have worked for the last three years, Chris Wilde has been a great help in taking care of all administration problems, and supplying me with coffee!

Several people in addition to those mentioned above have helped with the writing of this thesis. Brenda Satherley has gone over several chapters in detail, and helped me improve the clarity and accuracy of the writing. Peter Gough was a great help when he took me aside and gave me some guidelines on writing, and read some of my work. Fiona Mackay has also proof read some chapters.

I also acknowledge the support of Cardinal Community Laboratories and the Cot Death Division of the National Child Health Research Foundation.

Over the last year, my wife Katherine has been a great support, always interested and encouraging me to keep at it. The same is true of my many friends and flatmates, with whom I have had many wonderful times at home, in the mountains, or on the rivers—they continually encouraged me to keep up the work.

Finally, I thank my parents Margaret and Adrian, who throughout my years of study have given me continual love and support.

Preface

The research presented within this thesis is concerned with the development and application of engineering techniques in the area of medical research, a field that is termed *biomedical engineering*. The motivation for this research is the investigation of Sudden Infant Death Syndrome (SIDS), or Cot Death.

During 1985, Dr. Rodney Ford, a community paediatrician, became involved in SIDS research in Canterbury. Dr. Ford recognised the need for the application of engineering techniques to help study infant physiology and clinical aspects of SIDS. Thus, Dr. Ford, the late Professor Richard Bates, Dr. Richard Fright, and my first supervisor, Dr. Kathy Garden, began a collaborative research effort between the Christchurch Hospital and the Department of Electrical and Electronic Engineering at the University of Canterbury. To assist in the study of infant physiology, a masters project was undertaken in 1987, by Mr. Richard Dove, with the aim of constructing a system to collect and store infant physiological signals. The system has become known as *BabyLog* [Dove 1988]. Since then, various people have been involved in collecting and analysing these signals, and the group is now known as the BabyLog cot death research group. My research evolved as a part of this project.

In 1990, I became involved with the BabyLog cot death research group at Christchurch Hospital through a summer studentship. During this time, I developed software for the BabyLog system, and performed some initial tests on an algorithm for the detection of pauses in breathing during sleep, or apnoeas. Soon after the summer studentship, I began my Ph.D. studies. I extended the initial tests of the apnoea detection algorithm, and from there expanded my research to study apnoea detection in detail. Throughout this time, I have been involved as a member of the research team, which consists of several people with a variety of backgrounds.

When I started this research, my supervisor at the University of Canterbury was Dr. Kathy Garden. She encouraged me into many new areas, and was a constant support for my work. After she moved, Prof. Jim Reilly supervised me for six months, and he worked with me to devise new approaches to my research. I was then supervised by Dr. Jimmy Li for several years, and in partnership with him, I conducted the majority of my research. Finally, Dr. Phil Bones is my current supervisor in conjunction with Dr. Jimmy Li.

Dr. Jimmy Li has contributed to the engineering rigour in this research. He has provided many of the ideas that are the basis of the original research, and has constantly contributed to the work over the last four years. Throughout this thesis, wherever there is any original material, Dr. Jimmy Li has been a part of developing the research.

The structure of this thesis is outlined below, and the original aspects of my research are identified. Some areas of research have involved collaboration with other researchers, and these areas are also identified.

Note that English as opposed to American spelling conventions are used—for example, "apnoea" as opposed to "apnea."

Chapter 1 describes the context of this research, and introduces and provides the motivation for studying apnoea detection. Infant breathing and apnoea are described, and possible links to SIDS pointed out. From existing research, the objectives of this thesis are developed.

Chapter 2 explains how the data are recorded, and describes how sleep studies are performed on infants to gather physiological data. The BabyLog system is described, and the breathing signal that is used is discussed in detail. The characteristics that lead to difficulties that are later encountered in terms of accurate apnoea detection are well illustrated. The BabyLog system and signals have been described elsewhere, but the details of the interpretation of the breathing signal are new.

Chapter 3 investigates human expert interpretation of breathing signals and detection of apnoeas. The work was done in conjunction with expert clinicians who had extensive experience evaluating sleep study data. The experts detected the apnoeas, and developed definitions. I performed the analysis of the experts' results, and I also formalised their common interpretations of the breathing signal. The overall work in the chapter is original and was published in a medical journal [Macey et al. 1995].

Chapter 4 is a study of methods of analysing breathing signals. Performance measures for apnoea detection are developed, of which some are standard and one is new. A review of methods that are used for analysing breathing is included to illustrate other approaches to analysing breathing. The original research in the chapter consists of one performance measure, and the application of neural networks, peak-to-peak, and cepstral algorithms to apnoea detection. Previous research is clearly referenced.

Chapter 5 presents a statistical method of apnoea detection, and its application as part of the analysis of physiological data. The basis of the algorithm was originally developed by Dr. Richard Fright. My initial work consisted of testing and optimising the algorithm. I then developed a duration measurement algorithm to use in conjunction with the original apnoea detection algorithm. In the second part of this chapter, the algorithm is used as part of a study of normal babies in their home environment, led by Dr. David Tappin, the then Canterbury Cot Death Research Fellow. In consultation with Dr. Tappin and Ms Kerrie Nelson, a statistician, I developed algorithms for reducing the raw physiological data to a manageable size. Thus, I was involved in collaboration with others in the overall study, which was original research, and I specifically designed some new analysis techniques.

Chapter 6 presents a new method of apnoea detection, and is entirely original. The motivation and objectives for a new method are explained. The first part involves modeling apnoea signals by describing the signal properties that correspond to an apnoea. While the properties are original research, I relied on the help of Dr. Philip Schluter when designing the statistical tests. The second part of the chapter describes a system that uses the signal properties to detect apnoea. The system includes a standard neural network, but the application is original.

Finally, Chapter 7 concludes and suggests possible future research.

Following are some of the publications and presentations that have been prepared during the course of my Ph.D. research:

Ford, R. P. K., P. J. Brown, R. A. Dove, C. S. Tuffnell, and **P. M. Macey**, "HomeLog: long term recording of infant temperature, respiratory and cardiac signals in the home environment," *Journal of Paediatrics and Child Health*, **Suppl. 1**, pp. 26-33, 1992.

Ford, R. P. K., C. S. Tuffnell, **P. M. Macey**, T. M. Tappin, and M. Sambamoorthy, "Rectal temperature changes and apnea," Conference Program of the Fourth SIDS International Conference, p. 123, Washington, USA, June 23-26, 1996.

Macey, P. M., R. P. K. Ford, P. J. Brown, J. Larkin, R. W. Fright, and K. Garden, "Apnoea detection: human performance and reliability of a computer algorithm," *Acta Paediatrica*, vol. **84**, pp. 1103-1107, 1995.

Macey, P. M., R. P. K. Ford, and J. S. J. Li, "Reliable apnea detection from an abdominal breathing signal," Presented at: the Fourth SIDS International Conference, Washington, USA, June 23-26, 1996.

Macey, P. M., J. S. J. Li, and R. P. K. Ford, "Designing an expert system for apnoea detection," Proceedings of the Third New Zealand Conference of Postgraduate Students in Engineering and Technology, pp. 83-88, University of Canterbury, Christchurch, July 1-2 1996, 1996.

Macey, P. M., J. S. J. Li, and R. P. K. Ford, "Expert system for apnoea detection," *Engineering Applications of Artificial Intelligence*, accepted for publication, January 1998.

Macey, P. M., J. S. J. Li, and R. P. K. Ford, "Deterministic properties of apnoea in an abdominal breathing signal," *Medical and Biological Engineering and Computing*, under revision.

Tappin, D. M., R. P. Ford, K. P. Nelson, B. Price, **P. M. Macey**, R. Dove, J. Larkin, and B. Slade, "Breathing, sleep state, and rectal temperature oscillations," *Archives of Disease in Childhood*, vol. **74**, pp. 427-431, 1996.

Tappin, D. M., R. P. K. Ford, K. Nelson, B. Price, **P. M. Macey**, and R. Dove, "Central apnoea is not increased in normal infants after vaccination," Conference Program of The Fourth SIDS International Conference, p. 119, Washington, USA, June 23-26, 1996.

Tappin, D. M., R. P. K. Ford, K. Nelson, B. Price, **P. M. Macey**, and R. Dove, "The febrile stress of routine vaccination does not increase central apnoea in normal infants," *Acta Paediatrica*, vol. **86**, pp. 873-880, 1997.

In preparation:

Macey, P. M., R. P. K. Ford, and J. S. J. Li, "Comparison between two apnoea detection algorithms," for submission to *IEEE Transactions on Biomedical Engineering*.

Tappin, D. M., R. P. K. Ford, **P. M. Macey**, B. Price, K. P. Nelson, B. Slade, "Thermoregulation and the metabolic drive to breath in normal infants in the home during the first 6 months of life," for submission to *Acta Paediatrica*.

Table of Contents

Abbreviations

[.]	integral function (converts to integer)
$\cdot_{(j)}$	j^{th} order statistic
A/D	analog-to-digital
ALTE	Apparent Life-Threatening Event
bpm	beats per minute (heart rate)
ECG	electrocardiogram
FFT	Fast Fourier Transform
GB	giga byte
Hz	Hertz
MB	mega byte
MSE	Mean Square Error
REM	Rapid Eye Movement
SIDS	Sudden Infant Death Syndrome
SNR	Signal-to-Noise Ratio

Chapter 1

Introduction

Breathing implies life, and without breathing, life as we know it ceases to exist. All creatures that have ever walked on this planet have breathed, from the smallest insects to highly evolved mammals such as Homosapiens.

In humans, breathing functions are so fundamental to the animation of our body that the neurological control of breathing occurs at the subconscious level. Breathing, then, occurs instinctively without thought. People are seldom aware of their breathing and they breathe regardless of what they are doing. At the scene of an accident, upon noticing a body on the ground, a bystander might call out "Are they breathing?" to the same effect as "Are they alive?" This powerful association is again implied by the expression to *put a breath of life* into something. In some Eastern spiritual traditions, control of one's breathing is believed to be a path to spiritual enlightenment. Whichever way it is considered, breathing is very much associated with living.

During sleep, people breathe without conscious effort. Breathing continues throughout a night's sleep, consisting of regular, even breaths, although there may be periods of more erratic, uneven breaths as well. When sharing a room with another who is sleeping, one expects to hear slow but regular breaths. However, during sleep, people may be breathing regularly— and then they may pause for several seconds, for no apparent reason—and then resume breathing as before. Such a pause is not conscious, yet breathing may cease for five, ten or even thirty seconds. A pause in breathing during sleep is called an *apnoea* [Thach 1985].

This thesis is concerned with apnoeas. How and why apnoeas occur has been speculated on for years and has motivated considerable research, much of which is ongoing. Apnoeas are commonly studied from breathing signals, and although an apnoea is defined physiologically as a pause in breathing, there are few definitions of apnoea *signals*. This lack of definitions results in uncertainty about what signal shape corresponds to an apnoea, and hence leads to difficulties in accurately detecting apnoeas. The research presented within this thesis proposes new definitions of apnoeas in terms of a breathing signal, and also proposes new detection methods.

1.1 Infant Breathing

What is normal breathing? Is a five second apnoea normal? Is a thirty second apnoea normal? The answers to these questions are not simple, as there is great variation in what is considered normal breathing.

Breathing is controlled by the body according to a number of factors [Talbot and Gessner 1973]. Oxygen from the air is brought into the body cells, and carbon dioxide is removed from body cells into the air. This transfer of gases is achieved through the lungs and the circulatory system: as the ventilation of the lungs increases, the quantity of gas that is pumped into and out of

the circulating blood increases. The body cells alter the gas partial pressures in the blood as they consume oxygen and produce carbon dioxide. The rate of gas consumption and production is proportional to the cellular metabolic rate, which is dependent on factors such as physical activity, and heat production requirements. Chemoreceptors provide information on the oxygen and carbon dioxide partial pressures in the blood, and the carbon dioxide partial pressure in the cerebrospinal fluid of the brain. This information is conveyed to the respiratory control centre in the brain, which adjusts the ventilation of the lungs to maintain gas partial pressures at normal values.

At sea level, most people do not experience apnoeas, but above 6,000 metres regular apnoeas are common [West et al. 1986] At high altitude, one factor that cold contribute to the occurrence of apnoeas is the low level of carbon dioxide, as this could reduce the carbon dioxide partial pressure in the blood and hence reduce the breathing drive enough to allow pauses in ventilation. However, the majority of apnoeas are not caused by reduced levels of carbon dioxide. People with breathing disorders such as Sleep Apnoea Syndrome typically have many apnoeas during a night, regardless of altitude [Guilleminault et al. 1978]. There is considerable variation in people's breathing patterns, and what is a normal pattern of breathing is determined by an individual's physical state and environment.

Normal breathing for infants differs from that of adults. This difference is primarily due to infants' smaller size and elevated metabolic rate [Hill and Rahimtulla 1965], the different oxygen-carrying capacity of the blood [Delivoria-Papadopoulos et al. 1971], and less sensitive chemoreceptors [Wilkie et al. 1987]. Infants have apnoeas during sleep, a fact that is accepted as normal [Gibson 1996a]. In contrast, not all adults experience regular apnoeas. The term "sleep apnoea" in adults refers to a serious breathing disorder, where the breathing passages obstruct normal breathing [Guilleminault, et al. 1978]. Apnoeas from two to ten seconds in duration are common in infants, but apnoeas greater than 15 seconds are less frequent [Tappin et al. 1996b]. Apnoeas greater than 15 or 20 seconds in infants are considered abnormal [Kempe et al. 1974, National Institutes of Health 1987, Gibson 1996a]. An example of an apnoea is shown in Figure 1.1.

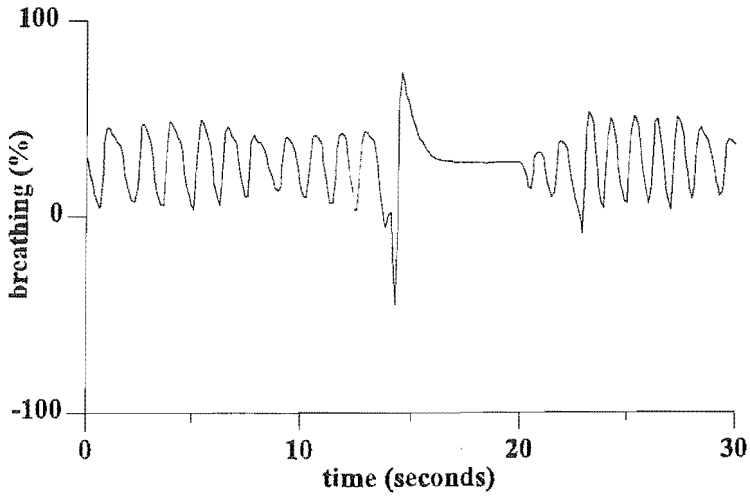

Figure 1.1 Central apnoea: a 30 second breathing signal with an apnoea starting at 15 seconds and ending at 20 seconds. The breathing signal is scaled relative to the maximum positive deviation and maximum negative deviation from the mid-range output of the instrument producing the signal, and thus the scale ranges from -100% to 100%. (For further details, see Section 2.3.3.)

While there is variation in the frequency and duration of apnoeas experienced by different infants, *all* infants have apnoeas during their sleep, with frequency ranging from less than ten to several hundred per night [Franks et al. 1977, Hoppenbrouwers et al. 1977, Stein et al. 1979, Richards et al. 1984, Hunt et al. 1985a, Henderson-Smart and Cohen 1986, Cornwell and Laxminarayan 1987, Lee et al. 1987, Tappin, et al. 1996b].

There are a variety of apnoea types. The most common is a *central* apnoea, where both breathing movements and airflow stop, often after a sigh, and then restart (see Figure 1.1) [Franks, et al. 1977, Stein, et al. 1979, Richards, et al. 1984, Hunt, et al. 1985a, Henderson-Smart and Cohen 1986, Cornwell and Laxminarayan 1987, Lee, et al. 1987, Tappin, et al. 1996b]. Sometimes, central apnoeas occur with such frequency that breathing only occurs for half the time. This phenomenon is called *periodic breathing* or *Cheyne-Stokes breathing* (see Figure 1.2). Periodic breathing occurs occasionally in infants, but not every night as occurs with central apnoeas [Kempe, et al. 1974, Franks, et al. 1977, Hunt et al. 1985b, Gordon et al. 1986, Southall et al. 1986, Gibson 1996a]. Another type of apnoea is termed *obstructive* apnoea, where airflow ceases but physical breathing movements continue, sometimes violently [Gibson 1996a]. While experiencing an obstructive apnoea, an infant behaves as if there is an obstruction in the airway [Brouillette et al. 1982, Brouillette et al. 1984, Guilleminault et al. 1984]. Obstructive apnoeas occur less frequently than central apnoeas and are considered abnormal [Brouillette, et al. 1982, Brouillette, et al. 1984, Guilleminault, et al. 1984, Dunne et al. 1986, Gibson 1996a]. It has been noted that some obstructive apnoeas are preceded by a central apnoea, and these have been labeled *mixed* apnoeas [Butcher-Puech et al. 1985, Gibson 1996a]. The purpose of this thesis is to investigate central apnoea in infants.

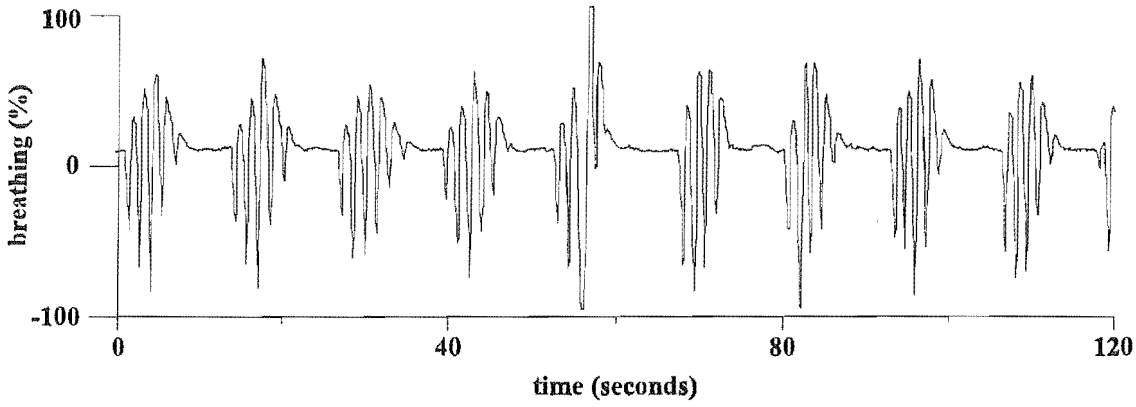

Figure 1.2 Periodic breathing: ten second apnoeas interspersed with ten second segments of breathing.

Superficially, central apnoeas appear to be unpredictable. Apnoeas occur more frequently in some infants than others. Infants with certain respiratory disorders tend to have more apnoeas and periodic breathing [Mitchell et al. 1983, Butcher-Puech, et al. 1985, Abreu e Silva et al. 1986a, Abreu e Silva et al. 1986b, Martin et al. 1986, Tudehope et al. 1986, Poets et al. 1993, Gibson 1996a]. However, according to recent research, most infants have at least ten central apnoeas of five seconds or longer during a night's sleep [Tappin, et al. 1996b].

Although the occurrence of apnoeas is unpredictable, there are physiological behaviours that are associated with apnoeas. After an apnoeic episode, the blood oxygen saturation may drop for a few seconds, reflecting the lack of oxygen being taken in by the lungs [Gibson 1996a]. A low

oxygen saturation is considered unhealthy, but central apnoeas do not appear to be harmful, with no recorded ill-effects [Gibson 1996a]. Other physiological behaviours have been associated with apnoeas, such as a change in heart rate [Haidmayer et al. 1982, Richards, et al. 1984, Butcher-Puech, et al. 1985, Kelly et al. 1986, Peirano et al. 1988, Kahn et al. 1992, Poets, et al. 1993, Barschdorff et al. 1994], but the precise mechanisms that trigger an infant to stop breathing remain unknown. Indeed, despite much research and many clinical evaluations, the meaning and significance of apnoeas are not completely understood.

1.2 Apnoea and SIDS

Every year throughout the world, thousands of babies aged between two weeks and twelve months die suddenly and unexpectedly [Gibson 1996b]. There is no apparent cause for such deaths; the baby simply dies during sleep. It is difficult to define this phenomenon, as there is no known cause, but it is commonly called Cot Death, or Crib Death in North America. Cot Death is specific to infants (four weeks to one year) as opposed to newborns (up to four weeks) or children (older than one year), and there is a peak in the death rate at three months of age. It is the leading cause of death for children under the age of 15 years in New Zealand and other developed countries. The medical term for this phenomenon is Sudden Infant Death Syndrome, or SIDS [Ford 1986, Hunt and Brouillette 1987, Nelson et al. 1989, Gibson 1996b].

During the 1980's, New Zealand had the highest recorded rate of SIDS in the world with a death rate close to 1% of births in some regions [Ford 1986, Nelson, et al. 1989]. This high rate motivated researchers to conduct an epidemiological study of all SIDS infants and a sample of control infants born in New Zealand over a three year period, resulting in new information on risk factors for SIDS. Three main modifiable risk factors were discovered: prone sleeping (sleeping face down); bottle feeding; and parents who smoke [Mitchell et al. 1991, Mitchell et al. 1992]. Publicising of these factors and uptake of different behaviours has substantially reduced the SIDS rate in New Zealand to average levels relative to other developed countries [Mitchell et al. 1994]. However, the study shed little light on the causes of SIDS.

Other research in New Zealand has focused on the study of infant physiology, searching for physiological mechanisms that could lead to SIDS [Tonkin et al. 1980, Bolton et al. 1993, Tuffnell 1993, Griggs et al. 1995, Macey, et al. 1995, Bolton et al. 1996, Ford et al. 1996, Tappin et al. 1996a, Tappin, et al. 1996b, Tonkin and Gunn 1996]. A hypothesis that is common throughout the world is that SIDS is caused by a failure of the respiratory system, and consequently much research focuses on breathing [Guilleminault et al. 1981, Hodgman et al. 1982, Gordon et al. 1984, Dunne, et al. 1986, Kelly, et al. 1986, Oren et al. 1986, Southall 1988, Milner and Ruggins 1989, Kahn, et al. 1992, Gibson 1996b, Katz-Salomon and Milerad 1996, Scheffer et al. 1996]. Research into breathing in New Zealand includes the use of physical models to simulate breathing whilst sleeping prone [Bolton, et al. 1993], studying the upper airway [Tonkin, et al. 1980, Tonkin and Gunn 1996], analysing mathematical models of respiration [Tuffnell 1993], and studying hours of recorded breathing data [Brown et al. 1992, Tappin, et al. 1996a].

The BabyLog cot death research group at Christchurch Hospital is active in the area of recording and interpreting physiological signals [Dove 1988, Brown, et al. 1992, Ford et al. 1992, Tuffnell 1993, Griggs, et al. 1995, Macey et al. 1996b, Tappin, et al. 1996a]. The BabyLog group is a

multi-disciplinary team of researchers, combining medical, engineering and statistical expertise. The BabyLog research is centered around hypotheses concerning the physiological control systems of infants, including breathing control and apnoea [Macey, et al. 1995].

An apnoea can be viewed in the context of previous and subsequent physical behaviour, and not as an isolated event; the fact that an apnoea occurred may not be as important as *how* or *when* it occurred. The respiration, temperature and cardiac control systems differ between infants and older children, and it could be the nature of these differences that makes infants susceptible to SIDS.

Although external causes for SIDS have been postulated for many years (for example, in New Zealand and the United Kingdom, the recent theory regarding the production of dangerous gases from mattresses [Sprott 1996]), so far no research has conclusively confirmed any external cause. Hence internal factors are likely to hold the key to the cause of SIDS. In particular, one hypothesis is that infants' respiratory and temperature control systems sometimes fail to cope with significant increases in temperature which, for example, might occur when an infant is trapped under layers of bedding [Ponsonby et al. 1992, Skadberg and Markestad 1996]. Several SIDS babies in Christchurch were discovered totally covered by bedding, and with high body temperatures recorded soon after death. Temperature control is closely related to respiratory control [Tuffnell 1993], and the mechanisms of apnoea may give clues to possible mechanisms of failure of temperature and respiratory controls.

There are indirect links between apnoea and SIDS. The age distribution of the incidence of SIDS matches the age distribution of infants that experience apnoea regularly. Since an infant can simply stop breathing for no apparent reason, could they also simply not resume breathing? Results are inconclusive; there is a higher risk of SIDS amongst infants who suffer obstructive apnoea, but there is no conclusive link between shorter central apnoeas and SIDS [Guilleminault, et al. 1984, Dunne, et al. 1986, Southall 1988, Kahn, et al. 1992, Gibson 1996a].

Apnoeas have been associated with SIDS in the language associated with infants. Some infants appear to almost die during sleep, but are revived. They are found not breathing, clammy-skinned and blue, and sometimes require full cardio-pulmonary resuscitation to be revived. These are termed Near-Miss Cot Deaths, or Apparent Life-Threatening Events (ALTE's); they are also called Apnoea of Infancy [Ariagno et al. 1983, Brooks 1992]. Infants who have suffered an ALTE are at greater risk of dying of SIDS [Oren, et al. 1986]. The parents of ALTE babies in New Zealand are given monitors to use on their babies during sleep. These monitors are called either Cot Death Monitors or Apnoea Monitors [Ford et al. 1994]. One fact is evident: if an apnoea is long enough, an infant will certainly die. As there is no diagnosis of "death-by-apnoea," if an infant experiences an apnoea and does not resume breathing, and consequently dies, then the cause of death would be unexplained, and therefore diagnosed as SIDS.

These possible connections between SIDS and apnoea have led to extensive research into the relationship between central apnoea and SIDS [Guilleminault et al. 1975, Hoppenbrouwers et al. 1980b, Hodgman, et al. 1982, Gordon, et al. 1984, Pfeiffer et al. 1984, Gordon, et al. 1986, Southall, et al. 1986, Ward et al. 1986, Hunt and Brouillette 1987, Kahn et al. 1988, Southall 1988, Wilson et al. 1988, Milner and Ruggins 1989, Schechtman et al. 1991, Kahn, et al. 1992]. Much international research aims first to understand the normal physical behaviour of infants, and then investigate

specific mechanisms which could lead to SIDS [Stein, et al. 1979, Hunt, et al. 1985b, Lee, et al. 1987]. Thus, much effort is being directed towards understanding apnoea, breathing and other related factors. While it is accepted that there is no direct link between short central apnoea and SIDS, research continues based upon the premise that the mechanisms of SIDS are related to the mechanisms of apnoea.

1.3 Detection Methods

The study of apnoea requires accurate apnoea detection. Detection is performed by analysing recorded breathing signals, as an apnoea is a pause in breathing. There is a variety of medical instruments that produce breathing signals, ranging from hand held monitors to gas partial pressure analysis systems installed at hospital bedsides. These instruments all measure some physiological behaviour associated with breathing, and they produce breathing signals that can be recorded electronically or printed. Recorded signals are usually interpreted by experts such as clinicians. In hospitals throughout the world, breathing is measured and recorded, and apnoeas are detected [Franks, et al. 1977, Hoppenbrouwers, et al. 1977, Stein, et al. 1979, Kelly et al. 1980, Guilleminault, et al. 1981, Haidmayer, et al. 1982, Mitchell, et al. 1983, Brouillette, et al. 1984, Pfeiffer, et al. 1984, Richards, et al. 1984, Butcher-Puech, et al. 1985, Hunt, et al. 1985a, Gordon, et al. 1986, Henderson-Smart and Cohen 1986, Kelly, et al. 1986, Martin, et al. 1986, Oren, et al. 1986, Southall, et al. 1986, Tudehope, et al. 1986, Ward, et al. 1986, Lee, et al. 1987, Kahn, et al. 1988, Peirano, et al. 1988, Milner and Ruggins 1989, Abraham et al. 1990, Kahn, et al. 1992, Poets, et al. 1993, Macey, et al. 1995, Ford, et al. 1996, Gibson 1996a, Tappin, et al. 1996b].

An overnight recording of breathing and other physiological variables during sleep is called a *sleep study* or *polysomnographic* recording. The field of *polysomnography* is relatively young, having developed over the last 30 years due to new instrumentation, recording and computing technology [Stein and Shannon 1975]. Polysomnographic systems have been introduced into hospitals to help diagnose patients such as preterm babies and ALTE infants [Dove et al. 1990]. Some systems have also been designed for home use in order to obtain research data [Franks, et al. 1977, Hunt, et al. 1985a, Gyulay et al. 1987, Ford, et al. 1992]. Systems for hospital use tend to record more signals than systems designed for home use, but the data from the home studies is considered valuable as the infant is in its normal environment.

A variety of signals are usually recorded for viewing and analysis. The particular signals recorded depend on the purpose of the study, and the protocol of the study centre. At Christchurch Hospital, infants with common problems such as gastro-oesophageal reflux may have only four signals recorded, for example oesophageal pH, breathing, blood oxygen saturation, and heart rate [Dove, et al. 1990]. Polysomnographic recordings for more complex cases, such as ALTE infants, usually include several breathing signals, several temperatures, two heart rate signals, blood oxygen saturation, and possibly other variables such as body position. Of all types of signals, breathing is the most commonly recorded, and apnoea detection is performed with almost all recordings of breathing signals.

Apnoea detection is used for different purposes. Apnoea *scoring* involves detecting all apnoeas greater than some prespecified duration, and giving a numerical score or frequency for each recording [Kendrick et al. 1990, Tappin, et al. 1996b, Tappin et al. 1997]. Another purpose of

apnoea detection is to locate longer apnoeas for viewing by a clinician. Routine hospital studies that record breathing for clinical evaluations use scoring, but concentrate on the apnoeas longer than 15 or 20 seconds, as these are considered clinically significant [Kempe, et al. 1974, National Institutes of Health 1987, Gibson 1996a]. Apnoea detection is also a built-in functionality of breathing monitors. If a long apnoea occurs, it is detected by the monitor and an alarm is triggered [Laxminarayan et al. 1983, Graseby Medical Ltd. 1988]. Apnoea detection for monitoring needs to be automatic with no expert verification; it does not matter if the breathing signal is slightly inaccurate provided that it allows successful detection of any long pauses. The research in this thesis, although relevant to all types of apnoea detection, is focused on the accurate scoring of apnoeas.

Having defined an apnoea physiologically as a pause in breathing, the question is: "How is an apnoea defined within a breathing signal?" Most breathing signals are not a direct measure of breathing in that they are not exact measures of the ventilation of the lungs, and therefore an apnoea signal must be an indirect measure of a pause in breathing. A signal showing a pause in breathing represents an apnoea, but what constitutes a signal shape that represents a pause in breathing? There is no agreed mathematical description of breathing or pauses in breathing, and so an apnoea is loosely defined by subjective human expert interpretation [Biernacka and Douglas 1993].

Traditionally, a trained expert looks through a printout of a night's breathing and records the apnoeas. As shown in Table 1.1, there is a variety of breathing signal types that are used to detect apnoeas. Many groups detect apnoeas manually, by viewing entire records of breathing [Kelly, et al. 1980, Douglas et al. 1982, Hunt, et al. 1985a, Lee, et al. 1987, Kahn, et al. 1988, Kahn, et al. 1992, Kirjavainen et al. 1996]. Some groups that record data onto computer still detect apnoeas manually [Richards, et al. 1984]. This is the original method dating from when polysomnographic signals were printed using chart recorders, before suitable data recording and storage devices were available [Stein and Shannon 1975]. Usually, all signals are printed out, and experts view these printouts, and search for apnoea patterns appearing simultaneously across several signals [Kahn, et al. 1988]. Manually detecting apnoeas by viewing entire recordings is a time consuming process. Depending on the display used, analysing one night's data can involve viewing well over 2000 screens of signals [Macey, et al. 1995]. Each apnoea that is detected has its duration measured, and the duration is recorded along with the time of occurrence, and possibly verified by at least one other expert. Trained experts may take from one to two hours to study a full night's recording of a single breathing signal [Macey, et al. 1995]. At present, there are no apnoea detection algorithms that emulate human expert detection.

Currently, some apnoea detection algorithms are being used, though usually in conjunction with a human expert [Burgess 1990, Biernacka and Douglas 1993]. An earlier detection system using analogue circuitry was considered suitable as an aid to a human expert, but not as an independent method of apnoea detection [Barnett et al. 1981]. Most commercial sleep study systems offer apnoea detection software that presents a variety of measures based on the number, time of occurrence, and duration of apnoeas [Burgess 1990]. Some research groups have their own detection software [Laxminarayan, et al. 1983, Wilson, et al. 1988, Macey, et al. 1995]. The software

Group	Transducer Type	Physical Behaviour	Output	Analysis
[Kirjavainen, et al. 1996]	-static-charge-sensitive bed	-movement	paper	manual
[Poets, et al. 1993]	-pressure capsule -inductance -nasal thermistors	-abdomen movement -chest & abdomen volume -nasal airflow	paper	manual
[Kahn, et al. 1988]	-abdominal & thoracic strain gauges -nasal & oral thermistors	-chest & abdomen movement -airflow	paper	manual
[Lee, et al. 1987]	-nose-piece & flow-through system -O_2 & CO_2 nasal analysers -$TcPO_2$ & O_2 at ear -rubber strain gauges	-inspiratory flow -airflow -$TcPO_2$ & O_2 saturations -chest & abdomen movement	paper	manual
[Beckerman and Wegmann 1985]	-microphone on chest	-breath sounds	tape	manual on oscilloscope
[Hunt, et al. 1985b]	-impedance electrodes	-chest volume	paper	manual
[Butcher-Puech, et al. 1985]	-abdominal strain gauge -nasal thermistor	-abdominal movement -upper nasal airflow	paper	manual
[Pfeiffer, et al. 1984]	-impedance electrodes	-chest volume	paper	manual
[Richards, et al. 1984]	-pressure capsule	-abdomen movement	tape to paper	manual
[Hodgman, et al. 1982]	-impedance electrodes -PCO_2 monitor -nasal thermistor	-chest & abdomen volume -expired CO_2 -airflow	computer	computer & manual
[Guilleminault, et al. 1981]	-abdominal & thoracic strain gauges -nasal & oral thermistors -nasal catheter -ear oximeter	-chest & abdomen movement -airflow -expired CO_2 -oxygen tension	paper	manual
[Franks, et al. 1977]	-microwave movement sensor	-body movement	tape	computer & manual
[Hoppenbrouwers, et al. 1977]	-impedance electrodes -infrared CO_2 pressure monitor -thermistor	-chest & abdomen volume -expired CO_2 -airflow	paper & tape	computer & manual
[Stein and Shannon 1975]	-impedance electrodes	-chest volume	tape	manual on oscilloscope

Table 1.1 Breathing signals used for apnoea detection by various research groups. The medium onto which the signals are recorded is shown, along with the analysis method.

usually has parameters that are set by the user, requiring some experience with the analysis (as explained by Burgess [1990]). Mason et al. [1974] presented a system that used a peak-to-peak measure to detect apnoeas, but no performance results were published. This method was found to be reliable for some signals [Hoppenbrouwers, et al. 1977]. Rakowski et al. [1986] developed an apnoea detection method based on the flatness of a signal. However, their system required manual input at the start of the analysis, and no performance figures were published [Rakowski et al. 1986]. A method based on detecting apnoeas from flat regions in the signal has reportedly been evaluated and used successfully [Bruckert et al. 1982], but results were not published. In the main, few commercial packages or customised analyses have published in depth details of their analysis algorithms and the performances of their algorithms, and so their accuracy and reliability is unknown.

Some groups have measured the performance of detection algorithms. A detection algorithm analysing an expired CO_2 breathing signal for long apnoeas, in a monitoring situation, was compared to a human scorer and found to have 1-2% false positives and 2-3% false negatives [Laxminarayan, et al. 1983]. However, an adult system using a breathing signal produced by impedance plethysmography (chest volume and abdomen movement) had sensitivity ranging from 94% down to 40%, with 18% to 36% of detections being false detections [Gyulay, et al. 1987].

Algorithms using different start and end definitions were compared, demonstrating that apnoea durations varied significantly, and apnoea density (apnoea seconds per hour) could double depending on the method used [Hunt et al. 1988]. BabyLog uses a detection algorithm but requires human expert verification [Ford, et al. 1992]. Biernacka has evaluated a detection algorithm and concluded that it was of little use [Biernacka and Douglas 1993]. Bruckert et al. [1982] presented detailed results of their algorithm compared to one expert, and found from 1% to 26% false negatives, and from 1% to 21% false positives. Several breathing signals including airflow were used, and it is unclear which signals the results were derived from. The performance figures are also difficult to compare against those of other groups, as the original false positives were re-presented to the expert who reclassified up to 17% of events.

Commercial systems rarely include figures on the accuracy of apnoea detection [Miles et al. 1989, Burgess 1990]. A study of three commonly used systems has been published, showing that even the most sophisticated apnoea analysis required that the users *...be highly trained in polysomnography and have at least 90% inter-rater reliability...*, and that 50 to 100 records needed to be manually scored to set analysis parameters, with ongoing validation once or twice per month [Burgess 1990]. A second system showed initial agreement between automatic apnoea analysis and individual scorer of 90%, but it was noted that *It is difficult to maintain this level of accuracy...* with frequent apnoeas [Burgess 1990]. The third system did not claim automatic apnoea detection, but called its analysis a "Scoring Assistant" [Burgess 1990]. Another system describes its classification of central apnoeas as "less successful" than analyses of other parameters [Miles, et al. 1989]. Thus, despite the fact that these systems recorded multiple signals, there were significant inaccuracies in the apnoea detection results.

There have been attempts to define breathing signal patterns corresponding to apnoea. Many definitions refer to physiological behaviour: a pause in breathing, the end of inspiration, or a resumption of breathing movements. Richards had one expert define events and train people to

analyse according to criteria that were not documented in the paper [Richards, et al. 1984]. Butcher-Puech defined the end of a central apnoea as the resumption of two or more breaths within three seconds [Butcher-Puech, et al. 1985]. A few signal definitions have been developed. Some groups that recorded more than one breathing signal defined central apnoea as occurring when more than one breathing signal was flat, but flat was never defined [Butcher-Puech, et al. 1985, Kahn, et al. 1992]. Rakowski et al. [1986] had an expert user set a threshold below which a signal was considered "straight" (the equivalent of flat), and an apnoea signal was defined as being straight for greater than a given duration. In another example, a flat signal was defined as being below a threshold of 25% of the previous breath [MacFadyen et al. 1988]. This definition could cause inconsistencies, as the relative amplitude of the previous breath ranges from a large, as with the sigh prior to the apnoea shown in Figure 1.1, to normal or small, as in Figure 1.2. A system designed for apnoea monitoring uses a slightly lower threshold of 15% of the previous breath [Laxminarayan, et al. 1983]. A peak-to-peak measure has been used, where an apnoea is defined as occurring where the time between two peaks is greater than a set duration [Mason et al. 1974, Revow et al. 1986]. An alternative approach has been used to test apnoea monitors. Rather than defining a signal mathematically, a series of training signals is used, effectively defining apnoea by example; the training signals can be either simulated or actual recorded data [Zoldac et al. 1993, Leverich et al. 1994]. Overall, apnoea definitions in terms of signal are rare, and almost no detection performance figures have been presented.

The problems of apnoea detection and the need for definitions of apnoea are well documented in the literature. In 1988, Southall reported such, stating: *The description of respiratory patterns is beset with the difficulties of definition...*, and *Clearly there is a need for accepted definitions of apnoea...* [Southall 1988]. Richards et al. [1984], in explaining their 21% observer variability, mentioned that: *The onset and end of a pause are not always easy to detect....* A conference on apnoea monitoring concluded that standard definitions for apnoea were urgently needed [National Institutes of Health 1987]. Miles et al. [1989] stated that apnoeas are difficult to distinguish, *...partly due to the fact that no major clinical organization has established unequivocal event definitions*. Kendrick et al. [1990] did a study on the scoring of apnoeas during sleep, and concluded: *(1) There is not complete agreement on the definitions of scoring apnoea, and (2) until such time as universal definitions are agreed, the precise definition of each type of apnoea should be included in epidemiological and intervention studies* [Kendrick, et al. 1990]. To quote a recent review, *A major problem in evaluating sleep studies is that there are no universal standards for scoring [apnoeas]* [Gibson 1996a]. Yount is more precise, explaining: *At the current time no group has defined the maximal amplitude and shape of signal change that will be declared to be a breath using each of the available transducing techniques, and as a result the absence of breathing effort [apnoea] is not established quantitatively* [Yount 1989].

1.4 Objectives of Research

At present, there are many problems associated with accurate apnoea detection, and these can be grouped into three categories. Firstly, there are few definitions of breathing signal shapes that correspond to an apnoea, and those that do exist tend to be subjective and open to a variety of

interpretations. Mathematical descriptions in particular are lacking. Secondly, the mathematical details of detection methods are rarely described, and descriptions that do exist usually require subjective interpretation. Thirdly, there are no standard reference sets of apnoeas or performance standards for apnoea detection. A lack of scientific rigour in the development of signal definitions and detection algorithms may be contributing towards the current lack of understanding of the exact mechanisms of apnoea. Therefore, the focus of this thesis is to improve apnoea definitions and apnoea detection methods.

Although it may be possible to use other breathing signals to increase the available information, this research is based on a single signal. Detection algorithms usually analyse one breathing signal [Laxminarayan et al. 1982, Burgess 1990, Macey, et al. 1995, Corwin et al. 1996], and experts often detect apnoeas using a single breathing signal, even though they may also use other types of signal such as oxygen saturation or heart beat to evaluate the severity of the event [Barschdorff, et al. 1994]. For the majority of home recordings, only one breathing signal is recorded [Ford, et al. 1992]. Therefore, any definitions or detection algorithms that aim to be widely applicable must apply to a single signal. Hence, the scope of this thesis is restricted to studying only one signal, and developing thorough definitions and detection algorithms for that one signal.

The goals of apnoea detection research include developing universal definitions that match human expert opinion of what signal shape constitutes an apnoea, and developing detection algorithms that detect apnoeas with a similar degree of accuracy to human experts. As a start towards these goals, the specific objectives of this research are as follows:

1. To construct a reference set of apnoea and non-apnoea events. Before definitions and detection methods can be developed, a reference is required. Human experts are the ultimate reference of what an apnoea is, and therefore the aim is to use human experts to develop a reference set of apnoea signals consisting of breathing signals and the apnoeas within those signals.
2. To develop a mathematical description of signal shape that accurately defines apnoea. A mathematical definition of apnoea within a breathing signal is a model of apnoea which can be adapted to a particular reference set of breathing signals and apnoeas. The aim is to eliminate any subjective reference and develop a completely objective description.
3. To develop performance measures for a detection system. A performance measure quantitatively describes how well a human expert or an apnoea detection algorithm performs, and is required in order to evaluate and compare detection algorithms. If groups used similar performance measures then results could be easily and appropriately compared. Each clinical result or research finding could be published with a reference to the accuracy of apnoea detection, and commercial apnoea detection software could include details of performance.
4. To develop a system that accurately and reliably detects apnoeas. Such a system would speed up analyses and improve the consistency and reliability of detection. The system would be based on the mathematical description outlined in the second objective.

5. To develop applications of apnoea detection, in the context of analysing physiological behaviour associated with apnoea. Apnoea detection fits within the context of studying physiological behaviour. Mathematical algorithms that precisely define the physiological analysis within which apnoea detection is performed would allow for accurate and rigorous analyses of physiological behaviour.

1.5 Summary

Breathing is common to all people but breathing patterns differ, especially between infants and adults. An apnoea is a pause in breathing during sleep, and infants have frequent apnoeas during a night. Apnoeas in infants are investigated, as it has been suggested that they are related to SIDS. In analysing apnoea, reliable detection methods and accurate definitions are needed, and these are the motivation for this research.

The research described here involves defining an apnoea within a breathing signal, initially by human expert detection, and then by developing a mathematical description. Based on the definition, a reliable detection system is to be developed. By developing mathematical definitions and detection algorithms with measured performances, this research aims to be a starting point towards developing definitions that match human expert opinion of an apnoea signal, and developing detection algorithms that perform with a similar accuracy to human experts. Ultimately, accurate apnoea detection would enable more accurate study of infant physiology.

Chapter 2

Recording and Analysing Breathing Signals

This chapter explains how breathing signals are recorded during overnight sleep studies. The recording system and breathing signal used in this research are presented, and the breathing signal characteristics are described in detail, along with the characteristics of the instrument that produces the signal.

2.1 Sleep Studies

Apnoea detection is performed on breathing signals recorded during overnight sleep studies, otherwise known as *polysomnographic* studies. Over the last 20 years, the field of polysomnography has developed into an important clinical area, with both research and clinical studies regularly performed in most major hospitals around the world [Parmalee et al. 1972, Stein and Shannon 1975, Franks, et al. 1977, Hoppenbrouwers, et al. 1977, Stein, et al. 1979, Guilleminault, et al. 1981, Douglas, et al. 1982, Haidmayer, et al. 1982, Hodgman, et al. 1982, Mitchell, et al. 1983, Richards, et al. 1984, Butcher-Puech, et al. 1985, Gordon, et al. 1986, Henderson-Smart and Cohen 1986, Kelly, et al. 1986, Southall, et al. 1986, Wilson, et al. 1988, Milner and Ruggins 1989, Abraham, et al. 1990, Ford, et al. 1992, Kahn, et al. 1992, Poets, et al. 1993, Gibson 1996a].

A polysomnographic study is defined as a recording of many signals (*poly-*) during sleep (*-somno-*), with the signals displayed as graphs either on paper or computer screen (*-graphic*) [Stein and Shannon 1975, Burgess 1990, Biernacka and Douglas 1993]. Typically, a variety of physiological signals are recorded overnight, while a patient sleeps. Examples of signals recorded include heart rate, breathing and temperatures. The current instrumentation, recording, and display technologies enable large volumes of data to be recorded and studied.

Sleep studies have been used to discover new information regarding infants' physical functions. This information relates to SIDS and other, less serious, infant disorders. There have been many studies comparing polysomnographic recordings of SIDS and non-SIDS infants, searching for differences in some physiological measure [Harper et al. 1978, Brooks 1982, Haidmayer, et al. 1982, Hodgman, et al. 1982, Guilleminault, et al. 1984, Pfeiffer, et al. 1984, Gordon, et al. 1986, Kelly, et al. 1986, Martin, et al. 1986, Oren, et al. 1986, Southall, et al. 1986, Peirano, et al. 1988, Wilson, et al. 1988, Schechtman, et al. 1991, Kahn, et al. 1992, Katz-Salomon and Milerad 1996, Scheffer, et al. 1996, Schluter et al. 1996, Tappin, et al. 1996a]. Sleep studies are also used as a tool to help clinicians make diagnoses [Dove 1988, Burgess 1990, Biernacka and Douglas 1993, Leverich, et al. 1994]. An example of a condition for which sleep studies are used is gastro-oesophageal reflux, where clinicians require precise information regarding the intensity, duration and frequency of reflux so that they can make informed decisions regarding treatment [Dove, et al. 1990, Vandenplas 1992]. Thus, sleep studies can be used to obtain specific physiological

information regarding an individual infant, and also general physiological information that relates to all infants.

The instrumentation, recording and display methods vary between polysomnographic systems, ranging from event recorders, which are essentially sophisticated monitors with the capacity to record two or three signals for a few minutes [Corwin, et al. 1996], to sleep laboratories with comprehensive data collection and display systems, including dozens of instruments [Lee, et al. 1987]. Different systems may be used to study similar physiological behaviours and hence record similar physiological signals, but there are many different instruments that can produce any one type of signal. Even if two polysomnographic systems use the same model of instrument, the recorded signal can differ between systems due to instrument calibration, the sensors used, and the manner in which the sensors are attached. Different systems also record data in a variety of ways, such as: printing out on paper [Stein and Shannon 1975, Kahn, et al. 1988]; storing average values [Wailoo et al. 1989]; or storing a fully sampled set of signals [Dove, et al. 1990]. The result is that data recorded with one system or instrument are likely to have different characteristics to data recorded with another system or another instrument.

Systems designed for adults are generally not suitable for infants [Stein and Shannon 1975]. Adults tend to be more cooperative than infants and tolerate more invasive monitoring, such as a mask over the face for measuring ventilation. Infants usually try to remove any sensor attached to their face so the majority of instruments that are used to measure infants' breathing measure chest or abdomen behaviour. There is less physical space on babies, with premature infants sometimes smaller than an adult's hand, and the number or size of sensors used for adult sleep studies may not be practical for infant studies. Some instruments have different sensors for adults and babies, but often other, child specific, instruments are used [Stein and Shannon 1975, Corometrics 1985, BOC 1986, Graseby Medical Ltd. 1988, Dove, et al. 1990]. Compared to adult studies, there are usually fewer signals recorded during infant sleep studies. The result is that infant studies tend to record fewer data and use less direct measures of physical behaviour than adult studies.

Although a variety of signals may be recorded, some types are more common than others. Along with circulation, respiration is a constant process that is essential to life and that reflects the health of a person. When diagnosing patients, clinicians evaluate respiration and circulation: the trade-mark of a doctor is a stethoscope, with which he or she listens to air flowing in and out of the chest, and to the heart beating. There are many instruments available to measure breathing and, during sleep studies, several breathing signals are often recorded in order to obtain a clear representation of breathing. In fact, breathing is recorded during almost all polysomnographic studies.

Measuring breathing has a variety of uses ranging from monitoring to investigating the function of breathing and other behaviour. When monitoring a patient, the purpose of measuring breathing is to check whether the patient is breathing or not, and the information required is essentially whether there is some breathing activity or not. When investigating physical functions, more detailed information is required than for monitoring, as patterns of breathing and how they relate to other physiological behaviours are of interest. The number and type of breathing signals recorded vary accordingly, from a single measure (for example, abdomen movement) to several measures (for example, airflow at the nose and mouth, chest volume, and abdomen movement).

As the recording and display technology becomes more affordable compared to ten or twenty years ago, polysomnographic systems are becoming common in most major cities, and for many types of presenting conditions, sleep studies are routinely performed [Burgess 1990].

2.2 The BabyLog System

This section describes BabyLog [Dove, et al. 1990], a polysomnographic system that is used to collect the data for the research presented in this thesis. The name "BabyLog" has also come to refer to the cot death research group based at Christchurch Hospital but, unless explicitly stated, within this thesis BabyLog refers to the polysomnographic system.

2.2.1 Overview

BabyLog is a sleep study system which originated from Christchurch Hospital, and which is currently used in several other New Zealand hospitals [Dove 1988, Dove, et al. 1990]. The system is specifically designed for monitoring babies. In essence, BabyLog produces and records a number of signals that are measures of physiological behaviours. The hardware and software are custom designed to meet the requirements of the clinicians and researchers performing the studies. Although BabyLog is a general purpose polysomnographic system, it is most commonly used to study babies with respiratory disorders, gastro-oesophageal reflux, and ALTE's. Studies are also done in the home using HomeLog, a portable version of BabyLog [Ford, et al. 1992]. Currently, home studies are mainly used for research purposes [Brown, et al. 1992, Tappin, et al. 1996a, Tappin, et al. 1997]. Home studies typically involve recording fewer signals than hospital studies, but the studies are performed over a longer period of time, usually every night for at least one week. All data presented in this thesis have been recorded using the BabyLog and HomeLog systems.

BabyLog includes hardware for data acquisition between instruments and an IBM compatible PC, and the software to sample and store the data. Signals produced by various instruments are passed through a 12 bit A/D converter and sampled at rates appropriate to the signals being recorded. Examples of signals and rates at which they are sampled include temperature at 1Hz, breathing and heart rate at 10Hz, and ECG at 100Hz, as shown in Table 2.1. Each night's recording usually lasts 12 to 16 hours and contains 2MB to 5MB of data (or about 15MB if an ECG signal has been recorded). Data from home studies are recorded in approximately weekly blocks that are usually between 10MB and 20MB in size. The recordings are stored on removable optical disks, allowing immediate access to all data. For each night's recording, a complete set of signals can be displayed and analysed .

BabyLog allows a range of signals to be recorded and displayed, as shown in Figure 2.1. As there are a variety of reasons for doing a BabyLog study, there are many physiological behaviours that can be recorded, as shown in Table 2.1. Breathing signals are common to all recordings, both in the hospital and the home. Of the instruments used to measure breathing, the Phillips Graseby MR10, hereafter referred to as the Graseby, is the most commonly used as it is the least invasive [Graseby Medical Ltd. 1988]. The sensor consists of a plastic capsule taped to the infant's abdomen, and no electrodes or electrode gel are required. Some types of signals are produced using invasive instrumentation, such as oesophageal pH which requires a probe down

Instrument	Signal	Physiological Behaviour	Sample Rate (Hz)	Home
Link (BabyLog instrument)	breathing paradoxical breathing ECG heart rate	chest and abdominal impedance (measures of volume)	10 10 100 1	No
BabyLog nasal & oral thermistors	breathing	airflow at nose and mouth	10	No
BabyLog temperature loom (temperature sensitive diodes)	temperatures	sites: auxilla, shin, rectal, anal, environment & forehead	1	Yes
Corometrics Neo-Trak 502	breathing ECG heart rate	chest impedance (measure of chest volume)	10 100 1	Yes No Yes
Graseby MR 10	breathing	abdominal movement	10	Yes
Mercury switch	position	rotation of 0°, 90°, 180° or 270° (on front, side or back)	1	No
Ohmeda Biox	blood O_2 saturation heart rate	absorption of red light by blood, at some extremity (finger, toe, ear)	1 1	No
pH probe (with BabyLog instrumentation)	oesophageal pH	pH in oesophagus	1	No

Table 2.1 BabyLog signals and the instruments that produce them are shown. The physiological behaviour being measured is described, along with the sampling rate and whether the signal can be recorded in the home environment using HomeLog.

the throat, and hence these invasive signals are only recorded when necessary [Dove, et al. 1990]. The Corometrics Neo-Trak 502, hereafter referred to as the Corometrics, is another instrument that produces a breathing signal [Corometrics 1985]. The Corometrics is routinely used in hospital studies, often in addition to the Graseby, and also produces a heart rate signal and ECG signal. However, the Corometrics requires expertise in attaching the sensors, and is more invasive than the Graseby with two sensors stuck with electrode gel on either side of the chest [Corometrics 1985]. During a BabyLog hospital study, typically at least one heart rate and one breathing signal are recorded in addition to other signals. Figure 2.1 shows a two minute segment of a typical recording that includes two breathing signals (Channel C and Channel F) and two heart rate signals (Channel B and Channel D).

The hardware and software has been designed and developed by the BabyLog group [Dove, et al. 1990, Ford, et al. 1992], and advantages of using a custom-designed system include a detailed knowledge of how the signals are produced, and exactly what physical behaviour they represent. The effects of filtering, noise and other inaccuracies are known, and these can be taken into account when evaluating the signals. Some commercial systems have limitations in these areas, with signals filtered to improve readability at the expense of accuracy [Leverich, et al. 1994].

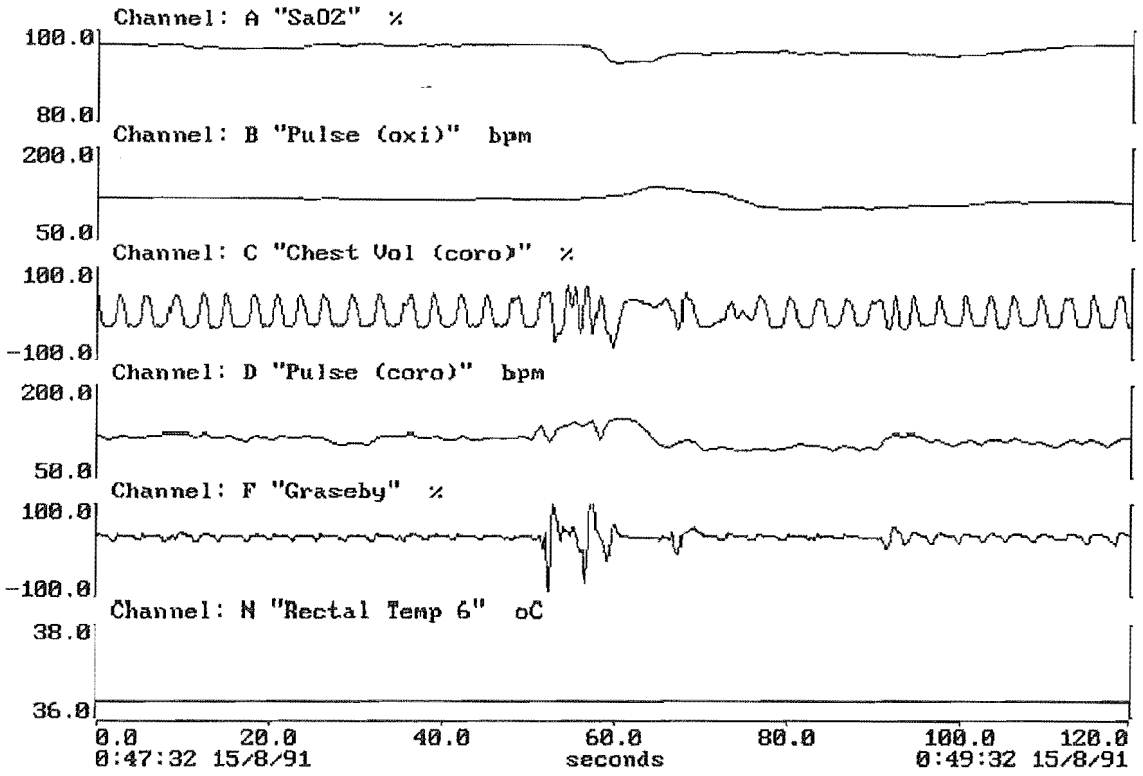

Figure 2.1 An example of a BabyLog display that illustrates six out of a total of eleven signals recorded during this particular study. Channel A is the percentage of oxygen saturation of the blood; Channel C is the Corometrics breathing signal; Channel D is the heart rate in beats per minute (bpm) as produced by the Corometrics; Channel F is Graseby breathing; and Channel N is rectal temperature in degrees Celsius. An apnoea event can be seen starting at 60 seconds, with the breathing signals (Channels C and F) being flat, the pulse (Channel B) increasing, and a desaturation event shown by the blood oxygen level (Channel A) decreasing.

Many commercial systems only allow limited adjustment of recording or analysis parameters, or may not describe algorithms used for automatic event detection [Burgess 1990]. Along with recording data, BabyLog has the capacity for viewing and analysing data, and displaying results. The data can be displayed in a variety of ways, and displays are often used to visually confirm calculated results. For example, simply displaying the recorded temperature data for a whole night has revealed the presence of rectal temperature oscillations, as seen in Figure 2.2, a phenomenon that had not previously been described in the literature [Brown, et al. 1992, Griggs, et al. 1995]. The BabyLog signals have minimal pre-filtering, and are as close as possible to the raw physiological data.

The standard signals recorded in the home are different to the standard signals recorded in the hospital. In hospital, sleep studies are performed for diagnostic purposes and the presenting condition dictates the number and type of signals recorded. An infant with gastro-oesophageal reflux might have recordings of oesophageal pH, Graseby or Corometrics breathing, heart rate, and oxygen saturation of the blood. A baby presenting with an ALTE would have recordings of at least two breathing signals, temperatures (rectal, anal, shin, abdomen and ambient), two heart rate signals, oxygen saturation, and body position (front, back or side). If breathing difficulties are suspected, three or four breathing signals may be recorded along with oxygen saturation and heart rate. In the home, infants are usually studied for research purposes, and the study is normally performed over several nights for up to seven weeks in a row [Brown, et al. 1992, Tappin,

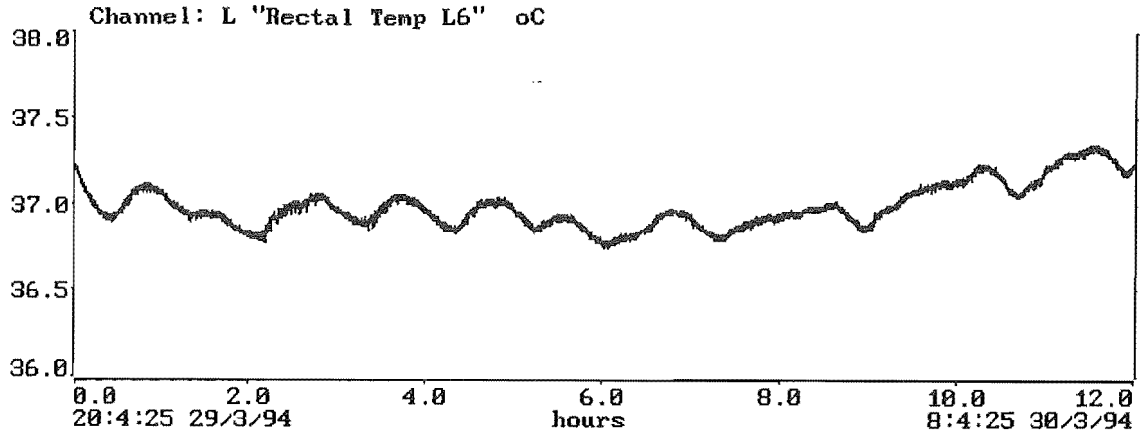

Figure 2.2 Temperature oscillations are illustrated, as seen in an overnight HomeLog recording of rectal temperature. Note also that the infant had 12 hours of uninterrupted sleep whilst the data was being recorded, a rare occurrence for BabyLog (hospital) studies.

et al. 1996a]. As part of various research projects, signals from each night's recordings are compared or combined, and therefore a characteristic of research studies is that the same signals are recorded each time. To minimise the disturbance to the child and to simplify the task of connecting the sensors, a small number of signals are recorded: Graseby breathing and temperatures. Hence, Graseby breathing is the only signal common to almost all BabyLog and HomeLog sleep studies.

The BabyLog system has been in use for ten years, and has been proven clinically. HomeLog has also been in home use for over six years, and there have been many publications based on the data recorded [Brown et al. 1990, Brown, et al. 1992, Tuffnell 1993, Macey, et al. 1995, Ford, et al. 1996, Macey et al. 1996a, Macey, et al. 1996b, Tappin, et al. 1996a, Tappin, et al. 1996b, Macey et al. 1997, Tappin, et al. 1997, Macey et al. 1998]. The main feature of the BabyLog system that distinguishes it from other polysomnographic systems is the accuracy of the signals in terms of the physical behaviour being measured, avoiding assumptions that lead to a blurring or reducing of the information contained within the signals.

2.2.2 Breathing Signals

As shown in Table 2.1, the BabyLog system includes four instruments to measure breathing:

1. Graseby (abdomen movement);
2. Corometrics (chest volume);
3. Thermistors (airflow at nose and mouth);
4. Link (chest volume & abdomen movement).

None of these signals is inherently better than the others, and they have their own advantages and disadvantages. As noted by Miles [1989]: *Non-invasive respiration recordings are often impaired by movement artifact, physical displacements of the sensor, and changes in body position....* More information can be available with more signals [MacFadyen, et al. 1988], and it has been shown that the detection of apnoeas can be improved with more signals [Upton et al. 1990]. There is also some evidence that more signals do not improve the accuracy of apnoea detection, but this was for the detection of apnoeas greater than 20 seconds, as opposed to apnoea scoring [Abdulhamid et al. 1992]. The four BabyLog respiration instruments allow a range of options depending on the purpose of the study being performed.

The Graseby measures abdomen movement using a pressure sensor taped to the abdomen, and is described in detail in Section 2.3. The Corometrics measures chest volume, based on the impedance between electrodes on either side of the chest [Corometrics 1985]. The thermistors are part of a custom designed and built instrument, and the sensors are placed on the infant's face. Sensors on the face are relatively invasive so thermistors are only used on infants with breathing problems, such as suspected obstructive apnoea or severe central apnoea. The fourth instrument is the Link four channel impedance system, a custom designed and built instrument that measures the impedance of the chest and of the abdomen. It measures the same physical behaviour as measured by the Corometrics, and in addition the abdomen movement is measured. The abdomen and chest movement signals are combined to give a measure of paradoxical breathing [Haidmayer et al. 1980]. With the example in Figure 2.3, the breathing changes from out of phase to in phase. The Link has only been in regular use since 1996, and so there are fewer recordings of Link signals than other breathing signals. The number of breathing signals recorded during a study varies; there is always one, and in the hospital often two, occasionally three as shown in Figure 2.4, and in rare circumstances, four.

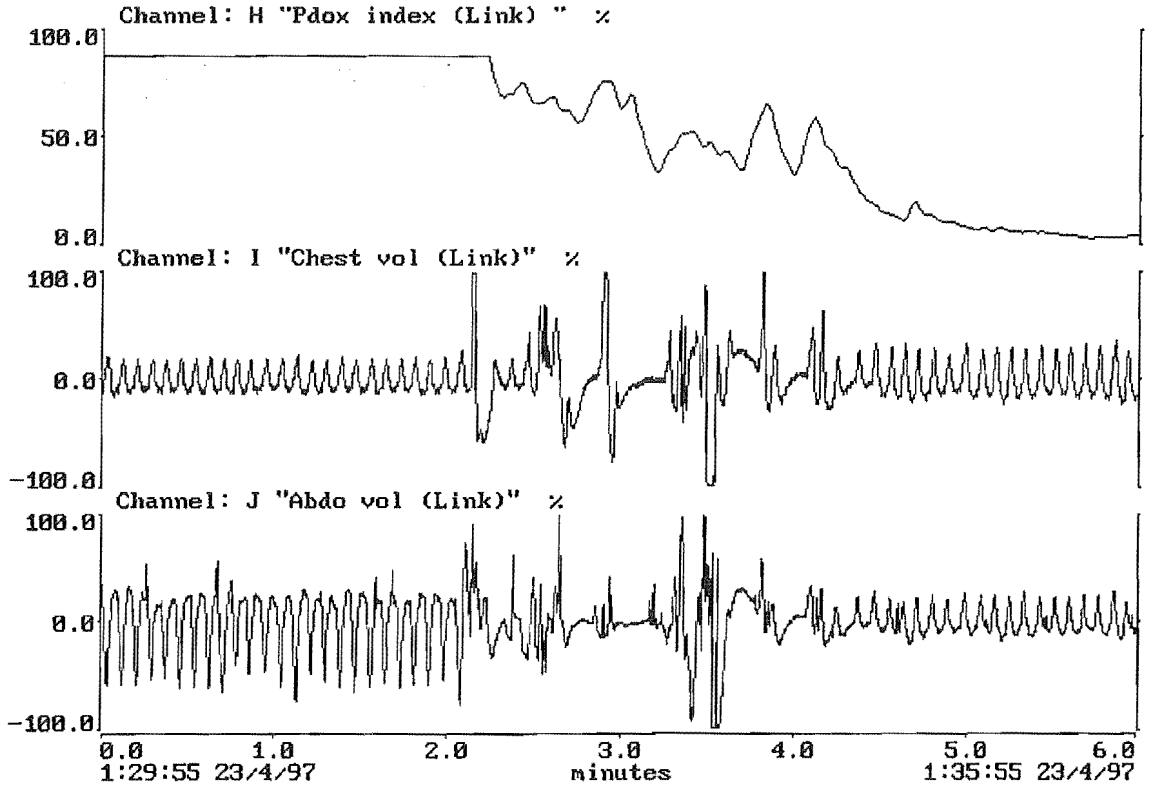

Figure 2.3 Link signals are shown, including the paradoxical index (Channel H) which is an estimate of the phase difference between the chest and abdomen. Note that a period of abnormal breathing occurs between 2 and 4 minutes, including an apnoea at 3 minutes, and the paradoxical index changes, indicating that the breathing changed from out of phase (paradoxical breathing) to in phase.

All breathing signals are designed to be a measure of true breathing, the ventilation of the lungs. The accuracy of each signal is different, depending on:

1. The accuracy of the physiological behaviour being measured compared to true breathing;

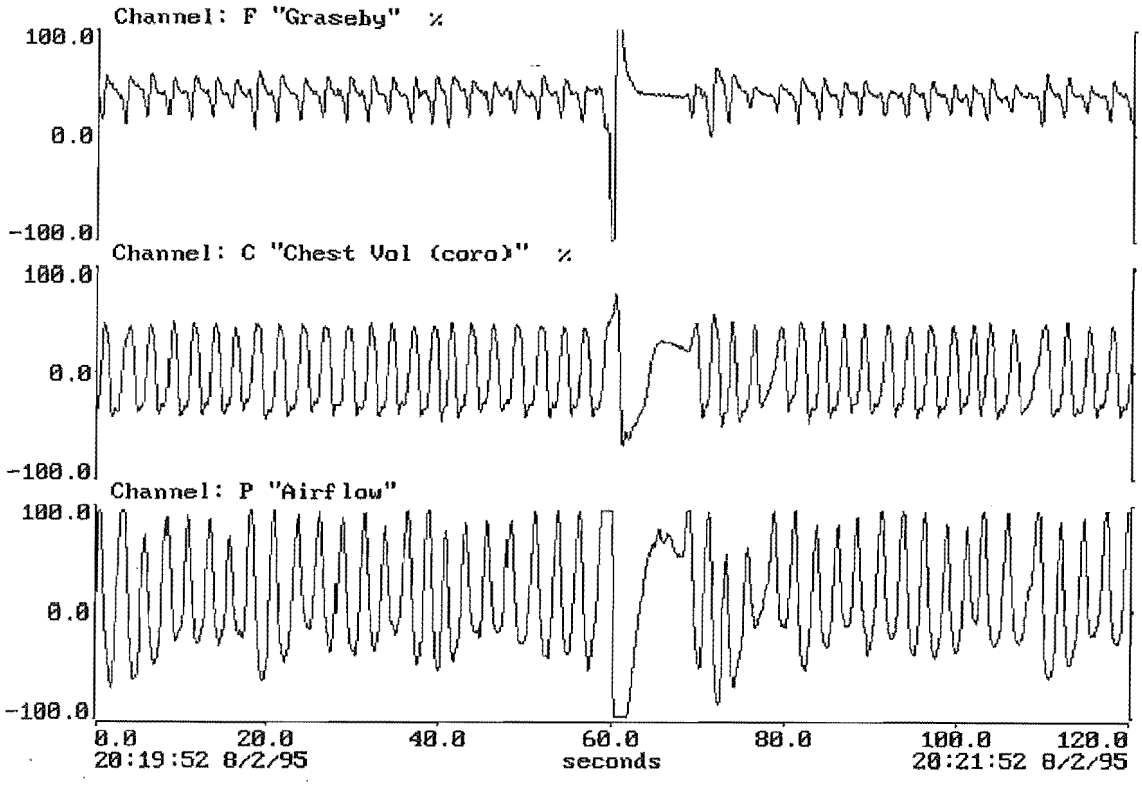

Figure 2.4 Three breathing signals recorded simultaneously during a BabyLog study, as seen on the BabyLog display. An apnoea is seen in the centre of the display at time 60 seconds, and is represented on all three signals.

2. Physical artifact, such as movement;
3. Instrument noise and filtering.

In practice, signals are produced by measuring some physical behaviour that is related to true breathing, but that is not a measure of the ventilation of the lungs. Given a physiological behaviour that is measured, the instruments and sensors used also influence the accuracy of the signal. If the instrumentation is invasive, the signal quality may be affected due to an infant being disturbed, and the infant either attempting to remove the sensor or sleeping restlessly. The sensors and instrumentation have some filtering or distortion, and therefore the signal is not an exact measure of the physiological input. The result is that the breathing signal characteristics are influenced by several factors.

Inaccurate or inconsistent signals occur for a variety of reasons. Most instruments require regular calibration, and therefore signals from two instruments of the same type are not directly comparable, even if they use the same sensor. Graseby monitors are recommended to be calibrated every six months [Graseby Medical Ltd. 1988], but in practice, calibration every three months was required to avoid excessive false alarms [Tappin, et al. 1996a]. Movement of the infant causes signal artifact, and signal noise can also be caused by sensors moving or being knocked [Mayotte et al. 1996]. Sensors may come loose, or have different characteristics depending on whether the infant is lying on them or not, or how securely the sensor is attached. Placement also effects the signal being measured, with some sensors, such as thermistors, needing exact placement. Over the last ten years of BabyLog studies, the most significant signal inaccuracies have been caused by poor sensor attachment, and baby movement.

To produce the BabyLog airflow signal, three thermistors are used. There is one below each nostril and a third just over the top lip. A single breathing signal is produced by combining all signals from all three thermistors [Dove 1988]. When the system is operating correctly, the thermistor signal is a relative measure of airflow at the nose and mouth. As airflow at the nose and mouth is more closely related to the ventilation of the lungs than abdomen or chest movement, the thermistors potentially produce the most accurate of the BabyLog breathing signals. However, the sensors are also the most susceptible to being incorrectly placed or being knocked after placement. With the BabyLog studies, the airflow signal has been the least reliable of all breathing signals, and in many cases many hours or even whole nights of recorded airflow data were just noise. An example of a poor airflow signal is shown at the bottom of Figure 2.6. To obtain a reliable airflow signal, the thermistors have to be placed such that the airflow from the nostrils *and* mouth passes over them. If the placement is slightly out, or maybe the infant has a blocked nose, the signal produced can be of poor quality. Infants, especially around the age of one year, tend to try and push the sensors off their face; the sensors can also get knocked as the infant moves. The airflow signal can provide useful information regarding some events such as obstructive apnoeas, but due to its unreliability, it is not considered a suitable signal with which to develop apnoea signal definitions and apnoea detection algorithms.

Another inaccuracy that has been reported by Storck et al. is that thermistors have a time lag, and therefore the timing of an airflow signal produced using thermistors is not accurate [Storck et al. 1996]. The timing of the breaths in the BabyLog airflow signal does not appear to differ significantly from the timing of the breaths in the Graseby or Corometrics signals, as seen in Figure 2.4, and the timing lag is not considered a significant source of error in this case. The small timing lag of the BabyLog thermistor is likely to be due to the small size of the thermistors, leading to a small thermal capacity and hence to rapid temperature changes [Dove 1988].

At Christchurch Hospital, a commonly used instrument to measure and monitor breathing is the Corometrics, as mentioned previously in Section 2.2.2 [Corometrics 1985]. The Corometrics is an impedance device, and in some countries, impedance devices are regularly used in hospital and home studies [Franks, et al. 1977, Stein, et al. 1979, Guilleminault, et al. 1981, Hodgman, et al. 1982, Pfeiffer, et al. 1984, Hunt, et al. 1985a, Lee, et al. 1987, Kahn, et al. 1992, Poets, et al. 1993]. In New Zealand, impedance monitors are seldom used in the home, because they are more complex to set up and more expensive than other monitors [Ford, et al. 1994]. The impedance electrodes (for the Corometrics and the Link) are smeared with electrode gel and securely taped to the body, with wires kept inside clothing and away from the face and hands. The Corometrics has alarms to warn of poor connections, but the electrodes tend to stay in place even when the child moves. The exact placement of the electrodes is not critical, as long as they are adjacent and on opposite sides of the chest. The impedance instruments are considered reliable monitors, and are also used to produce breathing and heart rate signals for the majority of BabyLog hospital studies [Dove, et al. 1990].

In the context of studying breathing, the BabyLog group has concerns with the accuracy of the Corometrics signal. The Corometrics filters the raw breathing signal extensively using an 11 pole adaptive filter, and tends to produce a breathing signal that is of a particular shape and amplitude, regardless of variations in the actual physical function during each breath. The

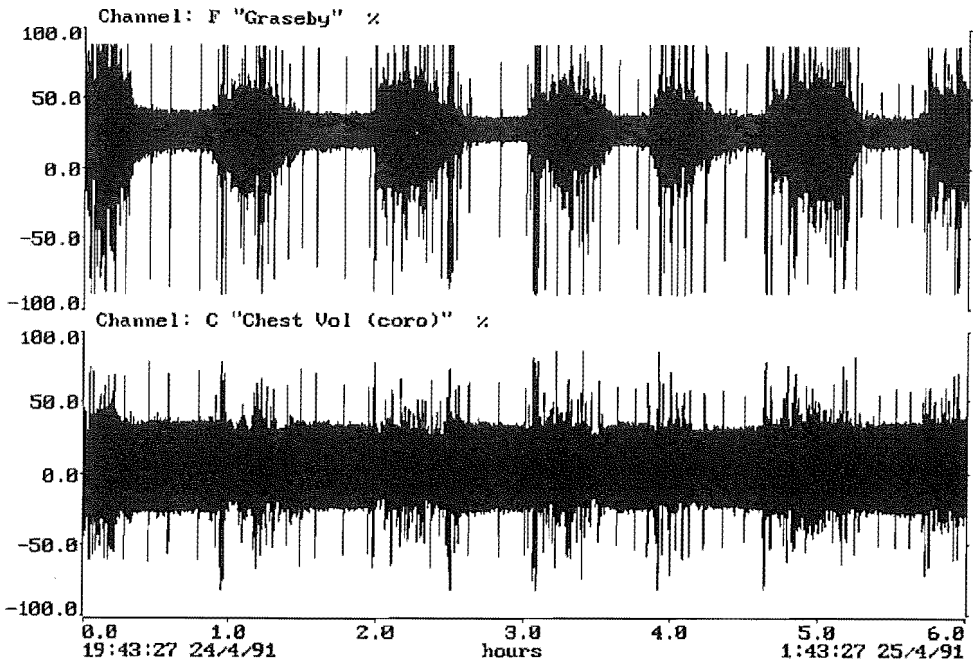

Figure 2.5 Corometrics and Graseby breathing signals are displayed for a period of six hours. Breathing state changes are more clearly reflected in the Graseby than the Corometrics signal, as the amplitude range of the Graseby signal changes more than the amplitude range of the Corometrics.

Corometrics signal appears as a clear breathing signal, but when compared with other breathing signals recorded at the same time, the Corometrics does not always reflect changes in breath amplitude, and sometimes displays normal amplitude breaths while other signals display shallow amplitude breaths or flat signals. It appears that the Corometrics filters the impedance signal to such an extent that it occasionally produces a signal with breath-like oscillations where experts agree no breathing took place. This inconsistency is illustrated in Figure 2.6, where the infant had only slight breathing movements, but the Corometrics produced a signal that contained oscillations of similar amplitude to normal breaths. An example over a period of hours is shown in Figure 2.5, where six hours of Corometrics breathing are compared with six hours of Graseby breathing; considering the envelope of the signals, the Graseby has greater variation between periods of low and high amplitude regions of signal, but the Corometrics varies little in amplitude range. Hence, the Corometrics is considered reliable as a monitor, but occasionally inaccurate in terms of producing a breathing signal for detailed analysis.

The Graseby is the least invasive BabyLog instrument that produces a breathing signal, with only one sensor on the abdomen [Graseby Medical Ltd. 1988]. The sensor is easy to place, as anywhere on the abdomen is suitable. The exact physical behaviour recorded may differ slightly depending on exactly where the sensor is located on the abdomen, and how tightly it is taped on. If no breathing signal is detected then the instrument will alarm, but there is no quality check of the signal. If the sensor is loosely attached, then the signal produced tends to remain at a low amplitude. Abdomen movement due to heart beat is less than in the chest, but it is still a source of noise. The Graseby is susceptible to movement artifact for two reasons: Firstly, the sensor is only taped on, and not glued with electrode gel like the impedance electrodes, and therefore the sensor can shift or loosen relatively easily; Secondly, as the infant moves the abdomen is likely to move, and hence movement is picked up that does not relate to breathing. The signal may also change

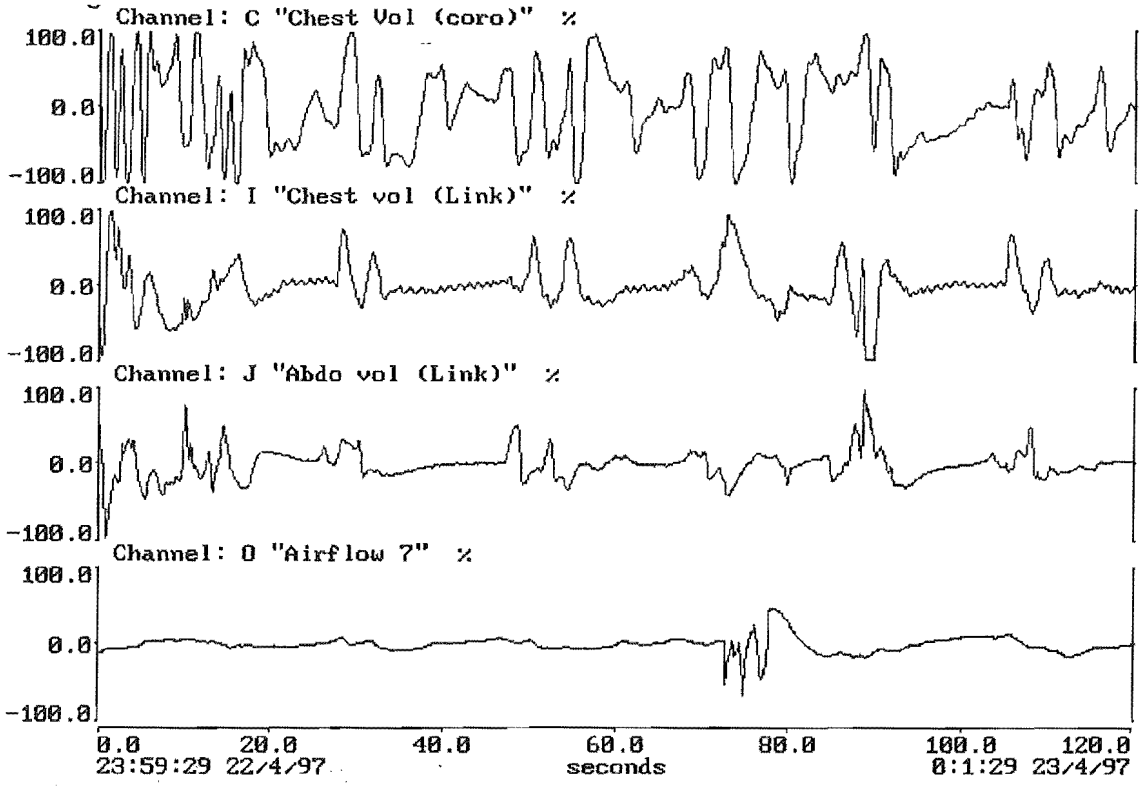

Figure 2.6 Example of a Corometrics signal (Channel C) showing oscillations that appear to represent breaths, but with chest and abdomen Link signals (Channels I and J) showing flat regions that experts confirmed as pauses in breathing. Experts judged that apnoeas occurred from approximately 35 to 45 seconds, and 55 to 65 seconds, but that the apnoeas were not accurately represented by the Corometrics signal. An apnoea also occurred from approximately 95 to 105 seconds, and is represented in both the Link and the Corometrics. Note also the poor airflow signal, Channel O.

dramatically after the child moves, especially if they roll onto or off the sensor. Further details of the Graseby signal are presented in Section 2.3.

Considering the four types of signal, the characteristics of each vary significantly. In Figure 2.4, the three signal types are indirect measures of the same behaviour, but they are not the same shape. If the accuracy of a breathing signal is defined as being "how closely it matches the ventilation of the lungs," no one signal type is accurate at all times, and each signal type has periods where other signal types reflect breathing more accurately. This lack of consistency is one reason why, where practical, multiple breathing signals are recorded.

2.2.3 Infant Database

The BabyLog and HomeLog studies over the last ten years have produced a large database of infant physiological signals. The database contains approximately 600 nights of hospital studies and close to 800 nights of home studies. Each study has at least one breathing signal, and most include a Graseby signal. In the past, some signals that are sampled at high rates, for example ECG at 100Hz, were discarded in order to save storage space. Nevertheless, the database consists of over 5GB of data.

The majority of the home data were collected during two research studies using HomeLog units [Brown, et al. 1992, Tappin, et al. 1996a]. A number of infants were studied over several

nights, up to seven weeks continuously. The infants were in their usual places of sleep, and had the sensors connected by their parents. Despite the environmental extremes of home conditions (such as cats urinating on computers and pet rabbits chewing connections!), most of the data are good quality. In contrast to infants studied in hospital, during home studies infants tend to sleep for most of the night, and hence the recordings are often free of movement artifact, and in general have much less awake time. Figure 2.2 illustrates a HomeLog recording gathered during twelve hours of uninterrupted sleep, which is a rare occurrence in the hospital studies. Hence, whilst home studies record fewer signals than hospital studies, the quality of the data tends to be better.

Advantages of hospital recordings include the variety and the number of signals recorded. The infants vary in age from premature babies through to four year old children, and they vary in health from healthy, low SIDS risk infants through to near-miss cot death victims. Although there are standard montages for common types of study, each study is designed specifically for a particular infant and its condition. When studying heart rate and breathing, several signals of the same type are often recorded simultaneously. As well as adding physiological information, this variety enables the characteristics of any particular signal to be better understood than if just one was recorded. Up to four breathing signals are recorded during any one study (see Figure 2.4 and Figure 2.6), which allows experts to obtain an accurate representation of the breathing behaviour of the infant. For the analysis of the home studies, where only one breathing signal is recorded, the experience gained during the hospital studies has helped the expert clinicians interpret the breathing signal.

The database is a resource which can be used to investigate infant physiology, and to study signal characteristics. New hypotheses and analyses are constantly being developed, and because the data are stored on removable disks, they are accessible and recordings are constantly being re-analysed [Tappin, et al. 1996a, Tappin, et al. 1996b]. At present, more data are being collected, and the database is a growing resource for testing analyses and investigating physiological patterns.

2.3 Graseby Breathing Signal

The Phillips Graseby MR10 respiration monitor is an instrument that is used for monitoring and for producing a breathing signal [Graseby Medical Ltd. 1988]. The Graseby is used throughout New Zealand for home monitoring, hospital monitoring, and sleep studies [Dove, et al. 1990, Ford, et al. 1992, Ford, et al. 1994, Macey, et al. 1995]. Similar abdominal wall movement measuring instruments are also used in other centres around the world for producing breathing signals [Butcher-Puech, et al. 1985, Gordon, et al. 1986, MacFadyen, et al. 1988, Schechtman et al. 1988, Abraham, et al. 1990, Hewertson et al. 1994]. The signal produced by the Graseby monitor is the breathing signal used in this research, and is herein referred to as a Graseby signal.

2.3.1 Selecting the Graseby

There are several reasons for choosing the Graseby. As a part of the objectives in Section 1.4, this work is to be as general and as widely applicable as possible, and therefore the breathing signal that is used to investigate apnoeas should be one that is widely used, and of a similar type to many other breathing signals. Most apnoea detection is performed on abdominal or chest breathing signals, as chest and abdominal breathing signals are the most commonly recorded

signals from infants (see Table 1.1) [National Institutes of Health 1987]. A definition and detection system based on a chest *or* abdominal signal would be widely applicable, as chest and abdominal signals are similar in nature [Brouillette et al. 1987]. Southall et al. [1986] compared the breathing signals from a pressure capsule and a jacket plethysmograph and respiratory inductance plethysmograph for 12 infants, and found that: *...the pressure capsule consistently detected the presence of each breathing movement.* The Graseby is an abdominal breathing signal, and is therefore of a similar type to the majority of breathing signals used for apnoea detection.

An advantage of choosing the Graseby signal is that, as mentioned in Section 2.3, the Graseby monitor is in common use throughout New Zealand and the world, both for monitoring and recording a breathing signal. In New Zealand, almost all home apnoea monitors used by parents of babies at high risk of SIDS are Graseby monitors, loaned out by the Cot Death Society, hospitals, and rented by pharmacies [Ford, et al. 1994]. In fact, the BabyLog group itself loans a pool of around thirty Graseby monitors to parents of babies at risk of SIDS [Ford, et al. 1994]. Some concerned parents even buy their own Graseby monitors. Hospitals in New Zealand use the Graseby for monitoring babies who have a low to moderate risk of SIDS. For patients at higher risk, the Graseby may be used in conjunction with other monitors, such as a Corometrics or a PulseOx pulse and blood oxygen saturation monitor (see Table 2.1) [BOC 1986]. The BabyLog and HomeLog systems use the Graseby for all research studies and almost all clinical studies. Thus, as well as being an example of a common *type* of breathing signal, the Graseby signal itself is in common use.

In the context of BabyLog, the Graseby has been in use for over ten years, and the BabyLog clinicians have extensive experience and knowledge regarding the signal. There is also a large database of Graseby recordings. Summarising, the Graseby signal was chosen because the instrument is widely used, the signal is of a common type, and there are many Graseby recordings in the database.

2.3.2 The Instrument

The Graseby MR10 is small, inexpensive and simple to operate compared to other monitors. It is one of the least invasive monitors available, and the sensor can be easily attached by parents. It has a high safety margin with no electrical connection to the baby and no mains power to the instrument. The instrument measures abdominal movement, producing a breathing signal for monitoring and recording purposes [Graseby Medical Ltd. 1988].

The Graseby consists of a case containing the electronics and battery, and a sensor, as illustrated in Figure 2.7. The sensor is a hollow capsule that is taped to the lower abdomen, and connected via a plastic tube to the case. The capsule responds to changes in the curvature of the abdominal wall. As the infant breathes, the abdomen moves, the pressure exerted on the capsule changes, and the air pressure inside the capsule changes, and a pneumatic signal is transmitted along the tube. The abdomen movements are correlated with actual inhalation and exhalation, therefore the pressure changes inside the capsule are related to breathing.

The sensor is semi-disposable: one sensor may last several months being used every night, but in general sensors require regular replacement due to small leaks caused by wear and tear. Parents who are loaned a monitor are typically supplied with three or four spare sensors, and many require extra over the period of monitoring (usually between three and twelve months) [Ford, et al. 1994]. The quality of the connection between the sensor and Graseby casing is as important as the quality of the sensor itself. The tube should be firmly inserted, and the seal must airtight. A faulty sensor or a poor connection leads to a weakened signal, which in turn leads to false alarms. Thus, parents are advised to replace the sensor if they are experiencing excessive false alarms. Fortunately, Graseby sensors are relatively inexpensive so regular replacement is not a problem.

There is a pressure transducer inside the instrument which produces an electrical signal from the pneumatic signal received from the sensor. Inside the case, the tube of the sensor feeds to the space between the plates of a capacitor, as shown in Figure 2.7. The capacitor has one fixed plate and one moving plate. As the air pressure changes in the capsule and tube, there is a pressure difference between the air between the plates and the air outside the capacitor. Thus, a force is exerted on the capacitor plates, and the moving plate shifts towards or away from the fixed plate, depending on whether the pressure of the pneumatic signal increased or decreased. The movement of the capacitor plate changes the capacitance of the capacitor, and the internal circuitry produces a signal proportional to the capacitance. The electrical signal is a measure of pressure changes that are in turn a measure of abdomen movement, and the signal produced is therefore an indirect measure of breathing. The instrument has an output from which the electrical signal may be read.

The capacitance of the capacitor is approximately 70pF, but varies both between different instruments and over time within the same instrument, from 55pF to 90pF [Graseby Medical Ltd. 1988]. This variation is the reason why regular calibration is required. Each instrument is

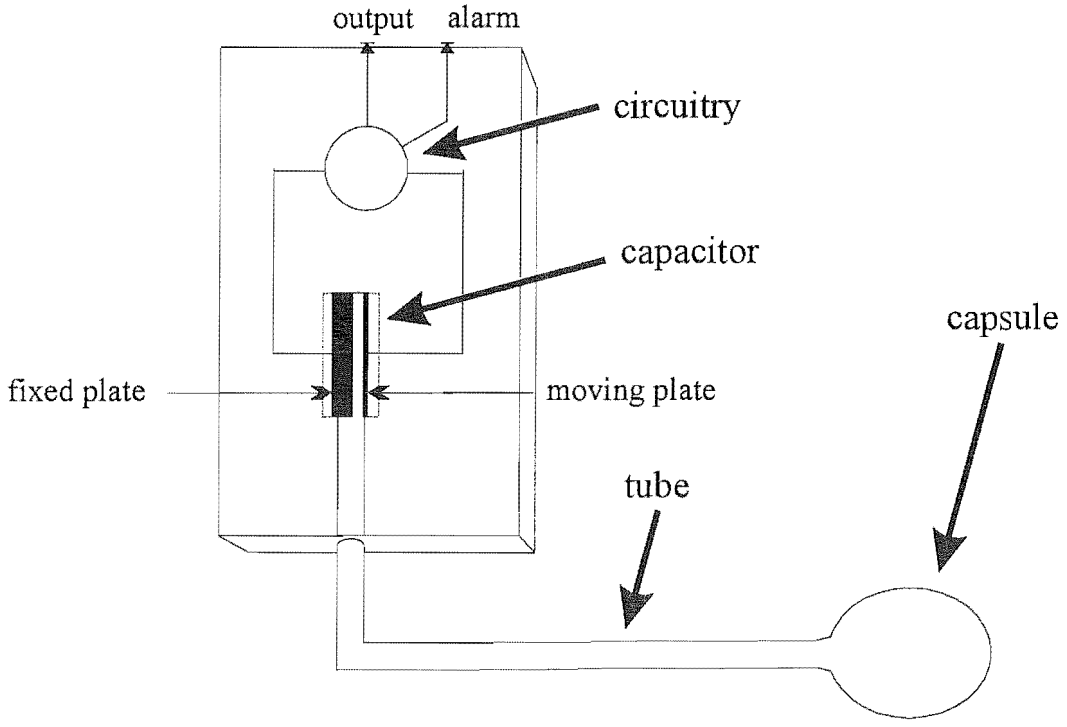

Figure 2.7 Simplified diagram of the Philips Graseby MR10. A signal is produced as a function of the capacitance, which is altered by the pressure signal from the capsule.

individually calibrated, and calibration is subjective, done by a human expert. Thus, irrespective of differences due to sensor characteristics, sensor placement, and noise, the Graseby signals may vary from recording to recording, and cannot be directly compared.

There is an alarm that activates if no breaths are detected within a certain period. The duration of this period is selected by the user. The monitor detects breaths with a threshold capacitor set to detect an increasing signal, corresponding to exhalation (assuming non-paradoxical breathing—see Section 2.3.3). As exhalation occurs, the signal increases, charging the threshold capacitor until the threshold is reached, and then a breath is detected. When a breath is detected, a light on the case flashes and a low volume click is produced, and the alarm timer is reset. Large and fast movements cause the threshold capacitor to charge faster, and therefore large breaths are detected earlier than shallow, small ones. Although the amplitude range of the signal may change, the threshold is fixed at an absolute value. To ensure *all* life-threatening apnoeas are detected, the threshold has a conservative setting, and the Graseby's breath detection is such that very few apnoeas are missed. This conservative setting is the reason that, although reliable as a monitor, the Graseby has many false alarms [Tappin, et al. 1996a, Tappin, et al. 1997].

2.3.3 Signal Characteristics

The Graseby breathing signal is a measure of abdomen movement, which is only an indication of breathing. A rising (expanding) of the abdomen is usually associated with inhalation, and a falling (contracting) abdomen is usually related to exhalation. However, this is not always so, as in the case of paradoxical breathing—a rising abdomen occurs with a falling chest and exhalation, and a falling abdomen corresponds to inhalation [Haidmayer, et al. 1980]. For the Graseby signals presented throughout this thesis, a rising signal usually corresponds to exhalation and a falling signal to inhalation.

The raw signal from the transducer is filtered to remove movement artifact, the DC component of the pneumatic signal, electrical interference, and cardiac artifact. The filtering is based on the assumption that breathing rate is between 40 and 140 breaths per minute (between 0.67 and 2.3Hz) [Graseby Medical Ltd. 1988]. The filter is bandpass with 3dB cutoffs at 0.5 and 3 Hz. As the normal range of heart beat for a newborn is 120 to 150 beats per minute (bpm), decreasing to an average of 100bpm by 12 months [Department of Health 1984], there is some frequency overlap between breathing and heart beat, and the resulting breathing signal may have noise due to cardiac movement [Bruckert, et al. 1982]. The filtered signal is sampled by the BabyLog system at 10Hz. A example of a typical Graseby breathing signal is shown in Figure 2.8.

As expected due to the overlap between possible breathing and heart rate frequencies, heart beat movement gets detected by the Graseby and is especially evident during breathing pauses. This characteristic is seen in other signals [Bruckert, et al. 1982]. During a breathing pause, cardiac movement usually appears as low amplitude oscillations. Gyulay et al. [1987] analysed an airflow signal and found detection errors occurred due small deflections of the signal during pauses in breathing. With the Graseby signal, the amplitude of the oscillations can reach the size of some breathing waveforms, a fact which has been noted previously [Southall et al. 1980].

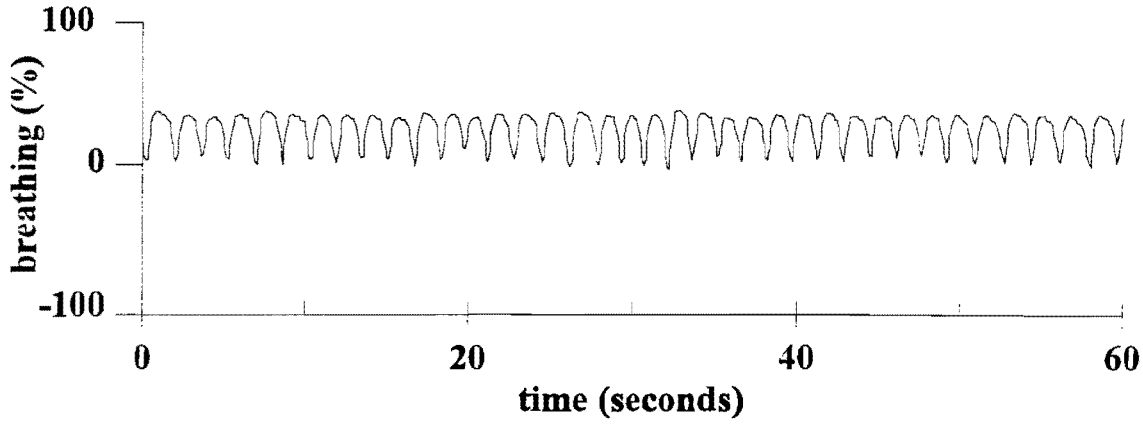

Figure 2.8 Regular breathing during quiet sleep recorded with Graseby.

Examples of such oscillations are seen in the regions R_{f1} and R_{f2} in Figure 2.9; both of these regions are considered by experts to be pauses in breathing.

The instrument has an automatic gain controller with a response time of approximately five seconds. The range of gain adjustment is limited so that small and large breaths are still represented by low and high amplitude signals respectively. In Figure 2.5, there are changes in the amplitude range of the Graseby, illustrating the fact that the automatic gain controller does not fully compensate for amplitude changes. In contrast, the Corometrics has significant filtering and gain control, as shown by the almost constant amplitude range of the recorded breathing signal in Figure 2.5 (see Section 2.2.2). Few details of the Graseby automatic gain controller are published, but overall it appears to have a minor effect on the signal [Graseby Medical Ltd. 1988].

The amplitude of the Graseby signal varies considerably and depends on factors such as the size of the abdomen movements, the position of the infant, and the reliability of the sensor attachment and tube connection. The average size of infants' breathing movements increases with age. For older patients, over twelve months, the breathing signal amplitude range is often high. For infants, under twelve months, high amplitude signals usually only occur when an infant is

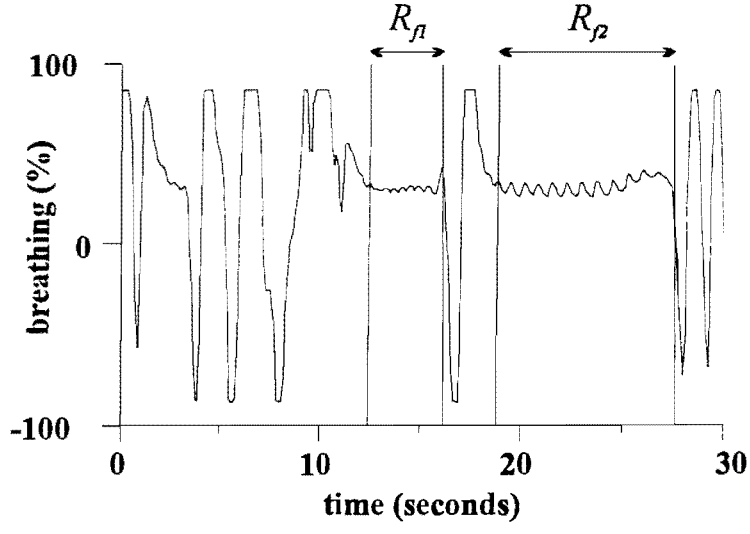

Figure 2.9 Cardiac oscillations with large absolute amplitudes are evident during the region R_{f2}, and smaller oscillations can be seen during region R_{f1}. According to the experts, both R_{f1} and R_{f2} correspond to pauses in breathing.

awake. The calibration of the instrument also makes a difference in the amplitude. When high amplitude signals occur, they can saturate the circuitry, causing a signal to oscillate from maximum to minimum values and hence be a distorted measure of breathing. This distortion is seen during the first 24 seconds of the signal in Figure 2.13, and to a lesser extent, with some of the breaths in Figure 2.9. Often the signal recorded by a Graseby monitor is relatively low in amplitude. The characteristics of a low amplitude signal are different to those of normal amplitude signals. A low amplitude signal is shown in Figure 2.10, illustrating how a significant proportion of the signal is flat, with oscillations corresponding to breaths having plateaus as opposed to the defined peaks of higher amplitude breathing signals, as seen in Figure 2.8.

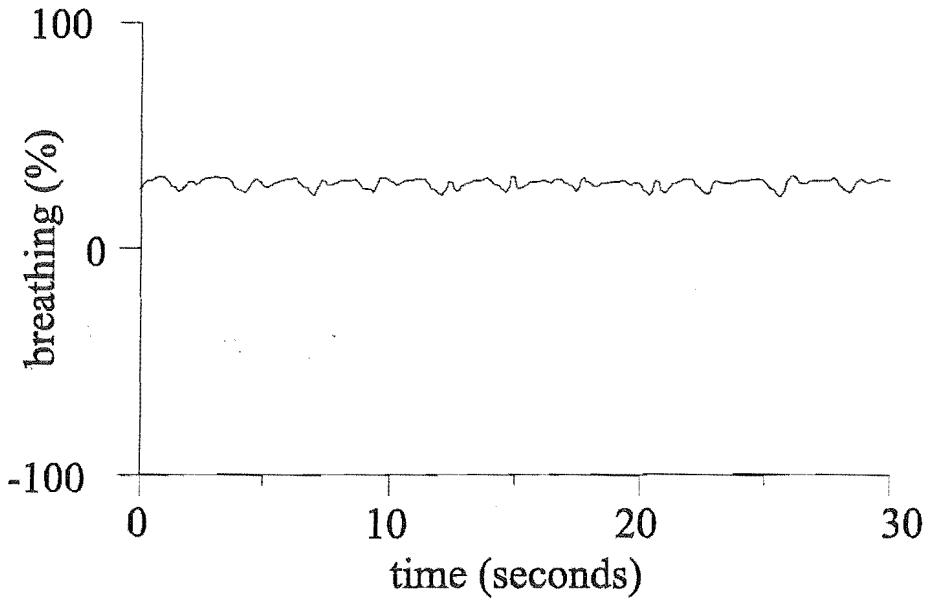

Figure 2.10 Low amplitude signal representing normal, regular breathing. Note how the signal is flattened during much of each oscillation, compared to a normal breathing signal as in Figure 2.8.

The relative amplitude of the Graseby signal can convey valuable information over a period of several minutes or hours. There are two types of breathing that can be distinguished, amongst others: regular breathing, as seen in Figure 2.8, and active breathing, as seen in Figure 2.11 [Tappin, et al. 1996a]. These types are relative to each other: regular breathing is more regular in breath amplitude and duration than active breathing. However, active breathing has less variation in breath amplitude and duration than breathing whilst the infant is awake, especially if the infant is crying or moving. Variable breathing that occurs while an infant is awake is more chaotic and high amplitude than active breathing. Regular breathing as in Figure 2.8 appears in a signal as peaks and troughs of similar amplitude, or as an approximately constant amplitude range over a long period of time, such as regions "Q" in Figure 2.12. Each breath oscillation also has a similar duration. Active breathing appears in a signal as peaks and troughs of varying amplitude, and of varying duration; these stand out over a whole night as regions with a higher envelope, such as the regions of signal other than regions "Q" in Figure 2.12. This difference in respiration has been noted by Mason et al. [1974]: *During quiet sleep, the respiratory waves are very regular and are of an even amplitude*; and *During REM [sleep] the amplitude and shape of the*

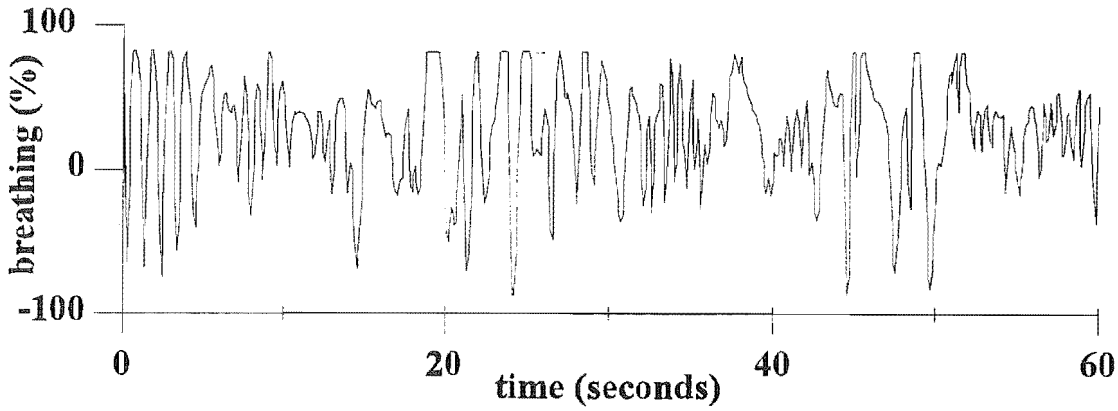

Figure 2.11 Active breathing as represented by the Graseby signal, with variable amplitude and duration of breaths.

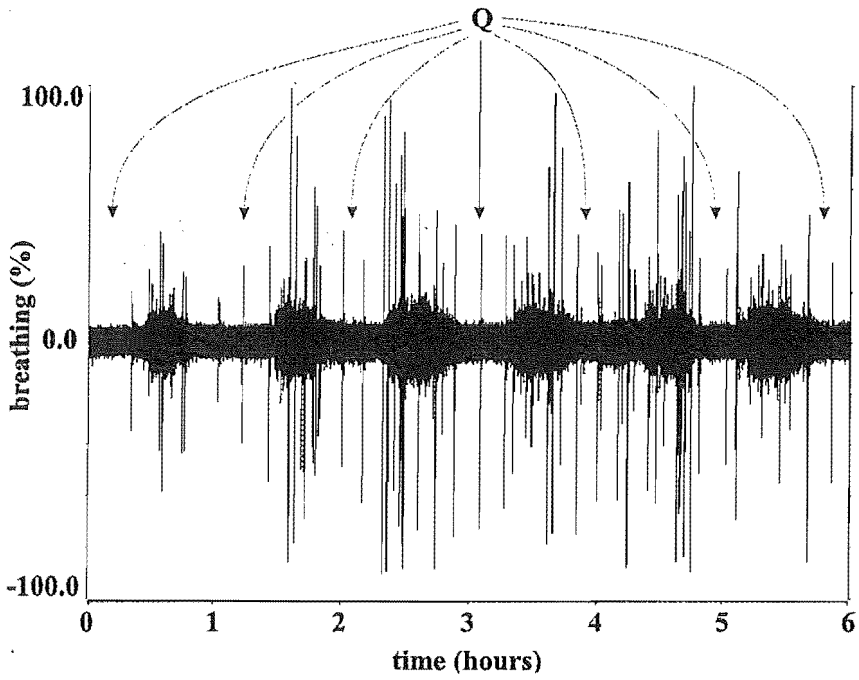

Figure 2.12 Six hours of Graseby breathing signal, containing in the order of 20,000 breaths. Note that the signal can be distinguish by the amplitude range, with the regions 'Q' having a small range compared to the remainder of the signal. Regions 'Q' correspond to shallow breathing during quiet sleep, and the remainder correspond to active breathing in REM sleep. The spikes are individual large amplitude breaths, and often correspond to sighs. (An example is the sigh prior to the apnoea in Figure 1.1.)

inspiratory and expiratory waves vary enormously.... A 30 second segment of active Graseby breathing is shown in Figure 2.11.

Breathing types relate to sleep states, such as Rapid Eye Movement (REM) sleep or quiet sleep [Wintrobe et al. 1974]. (There are other sleep states defined, but only these two are considered here.) Traditionally, sleep state is defined by a number of physiological measures including the electroencephalogram (EEG) and measures of eye movement, but these states can also be approximately defined using breathing signals [Harper et al. 1987]. Active breathing is associated with the REM sleep state, and regular breathing with the quiet sleep state [Harper, et al. 1987, Tappin, et al. 1996a]. When several hours of breathing are viewed in a single frame, as shown in Figure 2.12, the individual breaths are indistinguishable but the periods of regular and

active breathing, and hence quiet and REM sleep, are distinguished by the envelope of the signal (ignoring the spikes) [Tappin, et al. 1996a].

Movement of the infant can cause a large change in the amplitude of the signal, sometimes going from high to very low. The subsequent low amplitude signal can appear as no breathing, until the automatic gain control increases the amplitude, as in Figure 2.13. Experts distinguish these patterns as movements and not breathing pauses, but automatic apnoea detection systems often detect these low amplitude signals as apnoeas. While these changes in amplitude are relatively infrequent, with usually fewer than ten per night, they are a consistent source of false detections [Macey, et al. 1995].

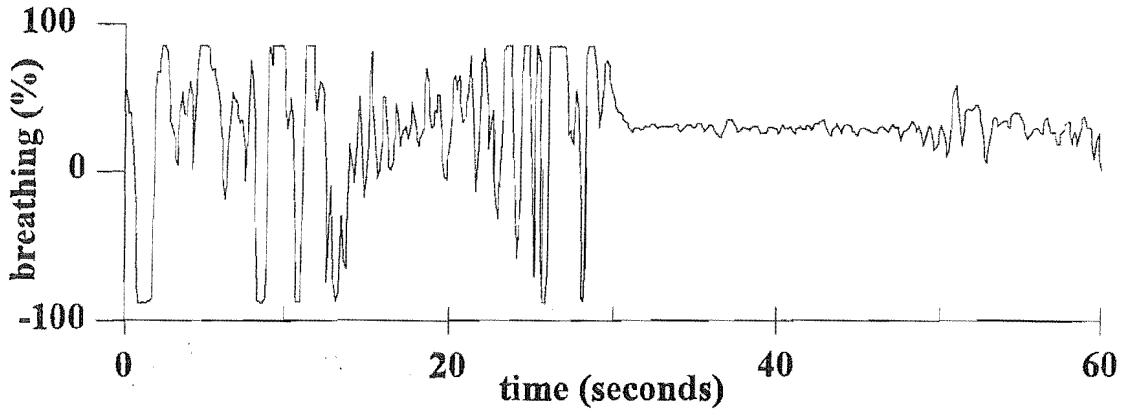

Figure 2.13 Graseby signal changing suddenly from high to low amplitude, probably caused by movement artifact. Note the saturation due to large breaths or movements, seen as flat peaks and troughs occurring between 0 and 30 seconds.

A final point is that the signal is a measure of pressure *changes* only. The pressure difference from between the inside and outside of the capacitor plates soon disappears through air seeping in or out. As the pressure increases, the signal decreases, but if the pressure stops increasing, the signal decays back to a rest value. Thus, even though there may be no breathing movement, the signal rises or falls. This characteristic is especially noticeable when studying apnoeas, and is discussed further in Chapter 3.

Summarising, breaths and pauses in breathing are in general reliably represented by the Graseby signal, both in terms of duration and amplitude. Changes in breath size are also reflected as changes in signal amplitude, and thereby different breathing states can be recognised by a visual inspection [Tappin, et al. 1996a]. The instrument itself is robust and safe. Overall, although the Graseby has some idiosyncrasies, it produces a reliable breathing signal.

2.4 Conclusions

Apnoea detection is performed on breathing signals, and breathing signals are recorded during sleep studies. BabyLog is a sleep study system that has been in clinical use for 10 years, and is well-proven for both clinical and research studies. A database of recorded signals exists of both hospital and home studies. In terms of investigating signals, the BabyLog database is a large resource of breathing data.

The objectives of this research centre around improving apnoea detection, and most apnoea detection is performed on chest or abdominal breathing signals. The signal recorded from the

Graseby monitor is used, as the Graseby is used extensively for both monitoring and producing breathing signals for recording during sleep studies. The Graseby is an indirect measure of true breathing, the ventilation of the lungs, and in general the signal is robust and contains the required information for apnoea detection.

The characteristics of the Graseby relate to the physical behaviour being measured, and to the instrument characteristics. The Graseby signal is an example of a common type of breathing signal that is recorded during polysomnographic studies, and is considered an appropriate signal to use for investigating apnoea signals.

Chapter 3

Human Expert Interpretation of Breathing Signals

In order to detect apnoeas from a breathing signal, a reference standard of the signal characteristics which constitute an apnoea is required. While the physiological definition of apnoea is a cessation of airflow, there is no standard definition of apnoea in terms of a breathing signal. Traditionally, human experts interpret a signal to determine what corresponds to a cessation of airflow, and whether the duration of the cessation is greater than some chosen minimum duration. Thus, the opinion of clinical experts is used as the gold standard for the particular shape of breathing signal that constitutes an apnoea.

Based on a number of overnight recordings, a set of apnoea and non-apnoea breathing signals is developed from expert assessment. This set constitutes a reference standard for developing definitions and detection algorithms. The agreement between experts is measured, and the expert interpretation of the breathing signals is studied.

3.1 Expert Detection of Apnoea

As explained in Section 1.3, the original method of apnoea detection involves a human expert looking through paper chart recordings of breathing and physically noting any apnoeas (Table 1.1). An expert is typically a clinician with experience in evaluating polysomnographic recordings. There are many experts around the world, with a variety of backgrounds and experience. Currently, each expert or group of experts formulates individual interpretations and definitions for apnoea detection. Consequently, different experts may detect different events and this subjective assessment of apnoeas is inherently imprecise. Nevertheless, expert opinion loosely defines apnoea.

There have been attempts to define signal characteristics that constitute an apnoea, but most of these definitions do not *objectively* define apnoea signals, and hence expert interpretation is needed. A pause in breathing is a simple concept and at a fundamental level experts use similar guidelines. However, many descriptions of apnoea signals refer to physiological behaviour. For example, the duration of an apnoea as measured from a breathing signal has been defined as the time from the end of expiration to the start of inhalation [Kahn, et al. 1992]; expiration and inhalation then need to be interpreted from the signal. Such a description is suitable for use by human experts, but not as part of an objective signal definition or detection algorithm. Probably the most common description of an apnoea signal shape is "flat" [Butcher-Puech, et al. 1985, Kahn, et al. 1992]. That a flat signal corresponds to an apnoea is not disputed, but describing exactly what is "flat," and where "flat" starts and ends, leads to differences of opinion, especially when taking into account the idiosyncrasies of any one signal.

Apnoea detection as has been performed to date is reviewed in Section 1.3. The key conclusion that can be drawn from this review of previous research is that there has been a lack

of consistency of apnoea detection. In particular, no clear signal definitions have been developed and no in depth comparisons between experts have been made. However, even though expert detection is subjective, existing methods have varying levels of objectivity. The approaches used by experts can be grouped into three categories according to the definitions used, as follows:

1. An individual's interpretation;
2. A previously developed description;
3. A mathematical definition of possible apnoeas, and these possible apnoeas then defined by either 1 or 2.

The first definition occurs when experts view breathing signals without considering other definitions, and interpret solely according to their own experience and knowledge. An example of using such a definition is described at the beginning of Section 3.2, where three experts detected apnoeas from a recording without any agreement or discussion about what signal shape represented an apnoea. The second definition occurs when an expert considers or uses a definition that has been previously developed, and is not just one person's interpretation. An example of the use of this second definition is also given in Section 3.2, where three experts agreed on a signal definition of apnoea and then detected apnoeas according to that definition, and finally discussed any disagreements. The third definition occurs where an algorithm is used to analyse a breathing signal over an entire recording in order to detect possible apnoeas for evaluation by a human expert. An example of this third method is presented in Chapter 5, where approximately 90,000 events that were possible apnoeas (flat regions) were detected using a software algorithm; experts then used the definition developed in Section 3.2.1 to classify these events as apnoea or non-apnoea. Thus, the definitions used range from totally subjective to a mixture of subjective and objective.

3.2 Reference Apnoeas

As mentioned in Section 1.4, an aim of this research is to develop a reference set of breathing signals corresponding to apnoea and non-apnoea. Such a reference set is required to test any definition of an apnoea signal, and as a standard against which to measure the performances of apnoea detection algorithms. As there is no universal definition of apnoea, any reference standard is relative to the opinion of experts. In order that this research be widely applicable, the definitions and algorithms are to be generally applicable, and therefore a reference standard would ideally represent the combined opinion of a number of experts.

A reference set of apnoeas has been developed based on the opinion of several experts. For the purposes of this research, a reference set is defined as a set of breathing signals and the set of all apnoeas that occur within the breathing signals, where the apnoeas are determined by one or more human experts. The method of developing a reference set was for an expert to examine a number of recordings and determine the epochs of signal that corresponded to apnoeas. Reference sets as determined by different experts were compared, and the consistency between experts was measured. A set was then created that represented the combined opinion of the experts, and a measure of expert consistency was associated with the reference set. As explained in Section 1.4,

an aim is to develop a detection algorithm that combines the expertise of many human experts so that it can perform better than one or more human experts.

Initially, a reference set was developed from one recording. Note that the whole recording was analysed, as opposed to rejecting regions of poor quality signal as has been done by other groups [Hoppenbrouwers, et al. 1980b, Kirjavainen, et al. 1996]. Three experts were used to detect apnoeas, each of whom had extensive experience analysing recordings and evaluating results, both with Graseby and other breathing signals. Three experts had also been used by another study in order to measure errors in scoring sleep studies [Bliwise et al. 1984]. These experts were all involved in the regular analysis of BabyLog studies. As BabyLog studies often record more than one breathing signal, the experts had experience at viewing Graseby signals alongside other signals, and were aware of the Graseby's idiosyncrasies as described in Section 2.3.3.

Once all three experts had completed a reference set for a particular recording, the agreement between the experts was determined. The results from each expert were compared and a measure of consistency calculated. Initially, it was thought that the experts would have good agreement and the main purpose of the expert apnoea detection was to obtain a set of reference events for testing the definitions and detection algorithms. However, in practice this was not the case, with initial results showing significant disagreement. The first recording analysed resulted in large discrepancies between the experts. All three experts had been instructed to list all the apnoeas longer than five seconds, without referring to the other experts; this is an example of the first type of definition as mentioned in Section 3.1. The first expert found 80 apnoeas, the second 112, and the third 149, and overall the experts disagreed on almost 50% of events. A similar high level of disagreement was also found with a study that measured the error in detecting apnoeas from adult breathing signals, using scorers with a range of experience [Bliwise, et al. 1984]. The variation demonstrated significant differences in expert interpretation of the signal. In order to develop a useful reference set of apnoeas, a more systematic approach was clearly required.

3.2.1 Definition

An agreement between the experts as to what signal characteristics constitute an apnoea was developed in the form of a definition. The experts formulated a definition that described all apnoeas within a Graseby signal, and the definition was discussed and refined until agreed upon by all experts.

A minimum apnoea length of five seconds was used. Clinically significant apnoeas are accepted to be pauses in breathing of more than 15 or 20 seconds, but for apnoea scoring, shorter events are also of interest [Kempe, et al. 1974, National Institutes of Health 1987, Brooks 1992, Gibson 1996a]. There is no agreed minimum length in the literature, but five seconds was chosen as a lower bound. Some groups use six or ten seconds as a minimum length, while others use only three or even two seconds [Guilleminault, et al. 1981, Hodgman, et al. 1982, Hunt, et al. 1985a, Kelly, et al. 1986, Southall, et al. 1986, Ghorbani and Bhavsar 1995]. If a two or three second threshold is used, it may be difficult to differentiate between a pause in breathing and a long breath, because some infants only breathe every two seconds. However, two to three second minimum durations are appropriate when studying babies with high breathing rates such as neonates or premature infants, for whom two seconds would be a definite pause in breathing. For the purposes of this

research, five seconds was selected as being sufficiently long to be a definite pause in infants' breathing. As well as being long enough to detect definite pauses, a five second minimum duration has the advantage of allowing for the detection of a greater number of apnoeas than would be detected using a longer minimum duration.

Although an absolute minimum duration is common, a variable minimum has also been used [Tappin, et al. 1997]. An apnoea has also been described as a cessation of breathing for a period of at least three times the median breath rate [Tappin, et al. 1997]. This period usually approximates to a duration in the order of five seconds, and typically ranges from three to eight seconds depending on the individual infant and its age. This measure, described in Chapter 5, has been used previously to calculate the breathing rate [Tappin, et al. 1996a, Tappin, et al. 1996b, Tappin, et al. 1997], and to define apnoea [Mason, et al. 1974]. The median breathing rate has been calculated over one to five minutes of breathing [Harper, et al. 1987, Tappin, et al. 1997], and also over a whole night [Mason, et al. 1974]. The median breathing rate can be calculated by taking the inverse of the median breath length, and breath lengths can be calculated using a peak-to-peak measure [Mason, et al. 1974, Revow, et al. 1986, Tuffnell 1993, Tappin, et al. 1996a]. There are a variety of other methods that can be used to measure breathing rate, each of which has advantages and disadvantages in terms of different types of signals (see Section 4.3) [Wilson et al. 1982, Moyles et al. 1989, Ning and Bronzino 1989, Wilks and English 1994]. The median filter is a robust operator in that extreme values have little effect on the output of the filter, and hence any errors which manifest themselves as outliers in breath length or peak-to-peak detection have minimal effect on the median breath length. However, there is no definitive and exact method of calculating the median breathing rate, as demonstrated by the variety of techniques available [Wilson, et al. 1982, Revow, et al. 1986, Moyles, et al. 1989, Ning and Bronzino 1989, Wilks and English 1994, Tappin, et al. 1996a]. In order to keep the definition and detection as objective as possible, a variable minimum is avoided, and an absolute length is used as a minimum apnoea duration for developing the reference set used in this study.

The three human experts agreed on a definition of apnoea for the Graseby signal taking into account its signal characteristics. They discussed their interpretations of the signal and the cues they used in recognising apnoeas. The experts then provided a description of the signal characteristics that they recognised as apnoea, in terms of a flat region and start and end points. The definition is as follows:

> *An apnoea indicated by a Graseby signal occurs where there is a flat region, usually preceded by a decay curve, and the combined length of the decay curve and flat region is at least five seconds* [Macey, et al. 1995].

Based on the preceding definition, an apnoea signal is essentially a flat region. The start of an apnoea is the first peak or trough prior to the flat region, and the end of the apnoea is the end of the flat region. A flat region is defined as being flat relative to surrounding breathing. However, in practice, this definition still requires a subjective interpretation of the signals, especially when the amplitude of the flat region approaches a significant proportion of the amplitude of the surrounding signals. The decay curve relates to a characteristic of the Graseby signal, which is that if there is no abdomen movement, the signal returns to a rest value, as explained in Section 2.3. The decay curve occurs in other signals, and is shown by Bruckert et al.

[1982] in airflow and thoracic breathing signals, who defined the start of an apnoea as the peak preceding a flat region. Thus, if an infant inhaled or exhaled and then stopped breathing, the signal would rise or fall to a peak or trough, and then at the time of the onset of the cessation of breathing, the signal would decay to the rest value. This characteristic can be seen in Figure 2.4, and is explained in more detail in Section 3.3. Deciding which peak or trough starts the decay curve, and exactly where the flat region ends, also require further interpretation. Nevertheless, having formulated a definition, the experts referred to the definition while detecting apnoeas, especially when classifying borderline events.

3.2.2 Methodology

In this section, the specific methodology of how apnoea detection was performed and how reference sets were constructed is explained.

Ten standard overnight BabyLog recordings from different patients were used. The choice of recordings was made without viewing the signals, and was based on the ten recordings performed within a one year period with standard montages. Ten recordings were used in order to include a variety of recordings and a large number of apnoeas. Recordings ranged from 12 to 16 hours in duration, and the patients were aged between three and 17 weeks, and all were full-term. Ethical approval for this study was granted by the Canterbury Medical Ethics Committee.

From the ten recordings, each expert detected all events that fitted the definition in Section 3.2.1. Each recording was viewed in its entirety, including the start and end regions when the babies were often awake. A standard approach was used: the signal was viewed in 30 second segments on the computer screen, shifting 20 seconds at a time, and only the Graseby signal was displayed. Flat regions corresponding to pauses in breathing were detected, and then the start and end points according to the definition were determined; the duration was measured to the nearest second. An event of a particular duration was defined as being of at least that duration so that a five second apnoea was at least five but less than six seconds long. Each apnoea start time and duration were recorded by the experts.

In detecting apnoeas, two aspects of performance that can be considered are sensitivity and specificity. *Sensitivity* is the percentage of apnoeas detected, and a high sensitivity means that few apnoeas are missed. *Specificity* is the percentage of detected events that are valid apnoeas, and a high specificity means that there are few false detections. The philosophy of apnoea detection is to initially detect all possible apnoea events and then evaluate these events further, and therefore a high sensitivity is important. This is in contrast to other analyses where specificity is more important, such as EEG spike detection [Gotman et al. 1978]. Although sensitivity cannot be measured, as there is no reference standard of what signal truly is an apnoea, the experts kept in mind the aim that *all* apnoeas were to be detected. If there was doubt with regards to classifying an event as apnoea or non-apnoea, the event was classified as an apnoea.

The method of duration measurement was not precise, but the accuracy of the duration measurement was not as important as the detection of the apnoeas. In other words, the accuracy of the duration measurement was secondary to obtaining high sensitivity. Therefore, a concern in terms of duration measurement was that an apnoea whose duration was close to the minimum was at risk of being not detected. The detected start or end times needed only a slight error for the

duration to be measured as shorter than the minimum duration, and hence be rejected as an apnoea. Therefore, if there was doubt regarding the exact time of the start or end point of a particular event, the start or end time was taken so as to maximise the duration. The trade-off was that some events that represented a pause in breathing of less than the minimum duration may have been be incorrectly detected as five second apnoeas, and also that some events may have been described as having a duration that was longer that the breathing pause they represented. Nevertheless, the sensitivity was of prime importance, and other performance measures could be compromised if the sensitivity was maximised.

Maximising the duration was therefore taken into account when detecting start points of apnoeas. The decay curve prior to a flat region could be movement or signal artifact, and therefore the true start of the potential apnoea cannot be determined from the signal (see Section 3.3). As mentioned above, it is more desirable to overestimate than to underestimate the duration, and therefore the decay curve was included in the experts' definition in Section 3.2.1.

After the experts had analysed the ten recordings, any disagreements on the identification of apnoeas were discussed among all three experts. This included disagreement about whether an event was an apnoea, and disagreement about the duration of an apnoea. The importance of reviewing events is illustrated by one study, where an expert reviewed events that he or she originally did not classify as apnoeas, and then reclassified up to 23% as apnoeas [Bruckert, et al. 1982]. As well as removing transcription errors and obvious mistakes, the discussion allowed experts to revise their decisions if they so desired. An apnoea list was then collated for each expert.

By comparing the apnoeas found by each expert from the ten recordings, the agreement between the experts was measured. Reference sets could then be compiled using the apnoeas that all experts agreed on, all apnoeas as detected by any expert, or other combinations.

3.2.3 Results

The numbers of apnoeas detected are presented in Table 3.1. Of the 619 apnoeas detected, 553 were detected by all three experts which is an average agreement rate of 90% (range 76% to 98%). Thus, there is a great improvement from the initial detection performed on patient 1, which had a disagreement rate of almost 50% (see Section 3.2). However, despite the experts having an agreed definition and discussing their differences, there remained an average 10% disagreement, and a range of 2% to 24% disagreement.

The duration of each apnoea was not always agreed upon: the experts usually agreed on the duration to within one second, as shown in Table 3.2. Of 553 apnoeas, the experts estimated the same length for 346 (63%), and 512 (93%) within one second. However, despite discussion, the experts disagreed on length by two seconds or more for 41 out of 553 apnoeas (7%). Therefore, of the 553 apnoeas that the experts agreed were apnoeas, 37% had duration measurements that the experts did not agree on. This level of disagreement reinforces the finding that there are significant differences in interpretation of where a pause in breathing starts and ends.

For the purpose of defining the duration of each apnoea, the longest estimate was used. The reason is that a high sensitivity is desirable, and taking the greater duration increases the likelihood of the apnoea being greater than the minimum duration. Most of the apnoeas are

Patient	Number of apnoeas detected			All Apnoeas	Agreed Apnoeas	Agreement
	Expert 1	*Expert 2*	*Expert 3*	*	†	‡
1	107	108	106	108	106	98%
2	66	68	72	72	66	91%
3	144	129	135	145	128	90%
4	95	95	87	97	87	90%
5	36	37	37	39	35	90%
6	35	33	34	36	32	90%
7	43	45	44	47	40	86%
8	13	14	13	14	12	86%
9	42	38	35	44	34	80%
10	13	17	16	17	13	76%
Total	**595**	**584**	**579**	**619**	**553**	**90 ± 6 %**

Table 3.1 Apnoeas detected by three experts from ten overnight BabyLog recordings are shown, along with human expert consistency.

* *All Apnoeas*: number of apnoeas identified by at least one expert;
† *Agreed Apnoeas*: number of apnoeas agreed on by three experts;
‡ *Agreement*: difference between *All Apnoeas* and *Agreed Apnoeas*, as a percentage of *All Apnoeas*.

relatively short, with 443 out of 619 (72%) being five or six seconds long (see Table 3.3). The longer apnoeas are rare, with only 17 out of 619 being ten seconds or longer. These findings are in line with other infant studies which have found many apnoeas less than ten seconds but few greater than 15 seconds [Stein, et al. 1979, Hodgman, et al. 1982, Southall, et al. 1986, Kahn, et al. 1988, Peirano, et al. 1988, Wilson, et al. 1988, Scheffer, et al. 1996, Tappin, et al. 1997].

Patient	Agreed apnoeas	Difference (seconds)		
		0	1	≥ 2
1	106	51	46	9
2	66	40	21	5
3	128	95	26	7
4	87	52	26	9
5	35	28	7	.
6	32	25	5	2
7	40	21	15	4
8	12	7	4	1
9	34	17	13	4
10	13	10	3	.
Total	**553**	**346**	**166**	**41**

Table 3.2 Consistency of length measurement between experts.

Most disagreement is over the shorter apnoeas: of the 66 apnoeas the experts did not agree on, 40 were five seconds in duration. Hence, the majority (59%) of the 10% discrepancy between the experts was due to disagreement about five second apnoeas, but the five second apnoeas only made up 45% of the total events. This difference suggests that there were events that all experts

agreed were pauses in breathing, but that some experts measured the duration of these pauses as greater than the minimum duration while other experts measured these pauses as being less than the minimum duration. This implies variations in the assigned start and end points of detected apnoeas. Of the 17 detected apnoeas that are ten seconds or longer, four were disagreed upon (24%); while this is a high rate of disagreement, the small number of events means that it is not statistically significant. Overall, there is a 14% disagreement over five second apnoeas, and an 8% disagreement over apnoeas six seconds and longer.

Patient	All Apnoeas	Duration (seconds)					
		5	6	7	8	9	≥ 10
1	108	43	39	15	7	2	2 (10,15)
2	72	31	22	8	6	3	2 (12,12)
3	145	80	30	20	7	3	5 (10,10,10,12,22)
4	97	40	13	14	18	8	4 (10,10,13,14)
5	39	18	14	5	1	.	1 (10)
6	36	22	8	4	1	.	1 (10)
7	47	22	14	7	2	2	.
8	14	1	6	5	2	.	.
9	44	11	15	7	5	4	2 (10,10)
10	17	11	3	1	1	1	.
Total	**619**	**279**	**164**	**86**	**50**	**23**	**17**

Table 3.3 Durations of apnoeas as measured by human experts. Where there was disagreement, the longer duration was taken.

Some recordings have greater disagreement, especially patients 9 and 10, and these differences of opinion appear to relate to the recording characteristics. When first seen, and before comparisons were made, the experts noted that recordings 9 and 10 were more difficult to interpret than the others. These recordings contained poor signals, often noisy and with frequent low amplitude segments, probably due to poor sensor placement. It was also noted that the recording for patient 6 is higher in amplitude than other recordings, resulting in unusual signal patterns. This high amplitude was probably due to the calibration of the instrument being out of normal range. Conversely, recording 1 had a breathing signal with consistent amplitude throughout the recording, and the events were clear and unambiguous, a fact that was reflected in the high agreement rate (98%).

The ten BabyLog recordings and the lists of apnoeas were combined to constitute reference sets that represented the opinion of all three experts. The ten recordings and the 619 apnoeas as detected by at least one expert are referred to as reference set B_1; the ten recordings with the 553 apnoeas as detected and agreed by all three experts are referred to as reference set B_{1ag}.

Summarising, there are two main aspects of the results. Firstly, human expert consistency has been measured, with a 10% disagreement between experts on what constitutes an apnoea, and a 37% rate of disagreement by one second or more on the duration. Secondly, reference sets that represent the combined opinion of human experts have been compiled, and can be later used for testing definitions and detection algorithms.

3.3 *Interpretation of Graseby Signal*

In detecting apnoeas, the experts had to interpret the Graseby signal. Based on an understanding of the signal characteristics described in Section 2.3.3 and of the physiology of infants, the experts evaluated the signal shape to determine whether it represented a cessation of breathing. Having detected a pause in breathing, the experts then interpreted the signal to determine exactly where the pause started and ended. In this section, some of the experts' interpretations and reasoning are explained.

An apnoea represented by a breathing signal is a flat region, but not all flat regions are apnoeas. The shape of the flat region is important in distinguishing between apnoea and non-apnoea. The flat region need not be totally flat in order to represent a pause in breathing, and may contain oscillations or unevenness due to cardiac oscillations or some other artifact, as explained in Section 2.3.3. However, an oscillation that is slightly larger than the typical cardiac oscillation, even if it is smaller than a normal breath oscillation, may be recognised as a breath by the experts. An example is in the flat region of the non-apnoea shown in Figure 3.1, where according to all three experts, point *a* corresponds to a slight breathing movement. Sometimes, a small breath may show up as an oscillation lasting a second or more, as seen at point *a* in the flat region of a non-apnoea event shown in Figure 3.2. In both Figures 3.1 and 3.2, the amplitude of the oscillation at point *a* is small compared to surrounding breathing, but large compared to the surrounding flat region. In classifying events as apnoea or not apnoea, human experts look at such details of the shape of the flat region in order to decide whether the signal represents breathing or no breathing.

Expert apnoea detection is subjective, but within any recording of breathing the majority of apnoeas are usually agreed upon. Disagreements arise over a set of events that are flat, but that could be classed as apnoea or non-apnoea; these events are referred to as *boundary* events. Many boundary events are not true pauses in breathing, but appear as a flattening of the signal, as explained in Section 2.3.3. These can be due to signal characteristics, a fact that has been noted previously: *respiratory pauses...were difficult to interpret since they occurred during periods lasting from 30 seconds to several minutes, in which large amplitude irregular signals occurred* [Jeffrey et al. 1981]. Jeffrey et al. also discovered false detections due to lower amplitude signals. In addition, there appear to be physiological events where infants almost stop breathing, as with the examples in Figures 3.1 and 3.2. The experts interpreted the signal as representing either breathing or a cessation of breathing based on years of experience of analysing such recordings.

Once a flat region of signal corresponding to a cessation of breathing was found, the duration of the pause was defined by the start and end points of the pause. The end of a breathing cessation is usually seen in the signal as a definite movement away from the rest value. After most apnoeas, the breathing restarts as normal, and there are few cases of gradual restarting of

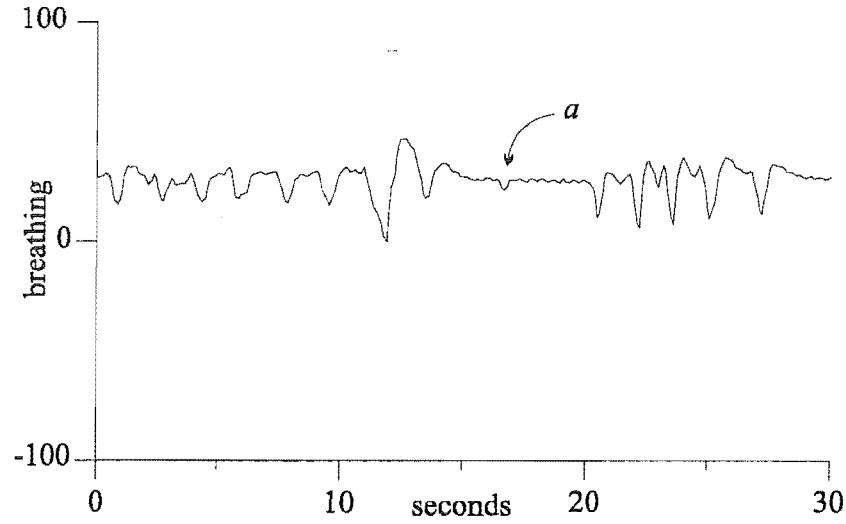

Figure 3.1 Non-apnoea event showing a flat region with cardiac oscillations, and a larger oscillation *a* corresponding to a small breathing movement.

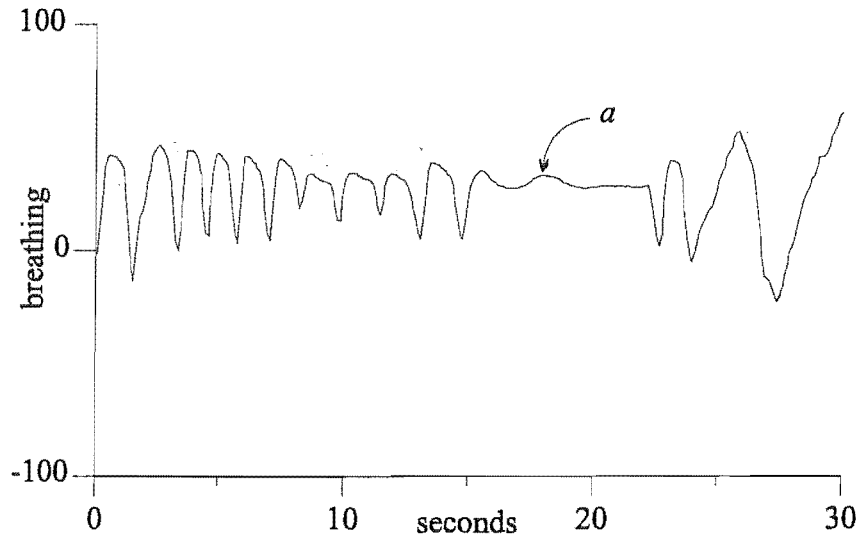

Figure 3.2 Non-apnoea event showing a flat region with oscillation *a* corresponding to a breathing movement.

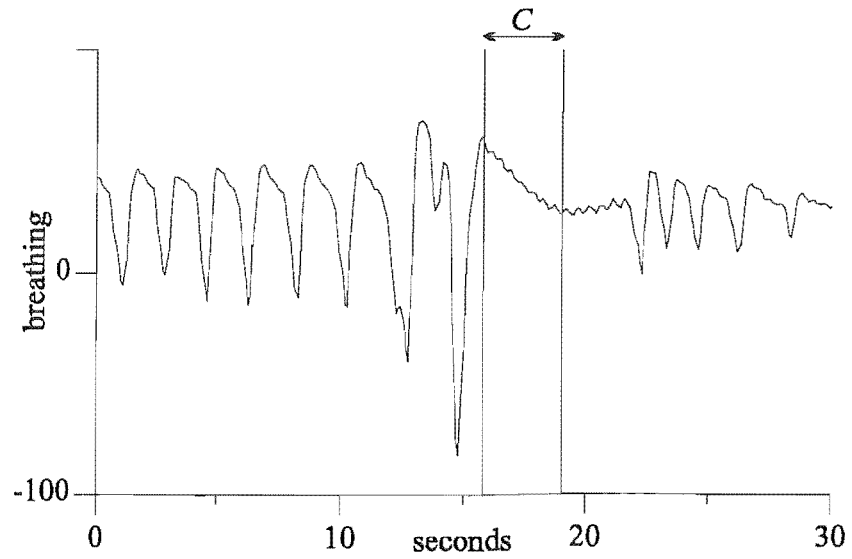

Figure 3.3 Apnoea with a decay curve *C* lasting approximately three seconds.

breathing, and even fewer where the infant gasps for air—takes a large breath. Comparing the ends and starts of the flat regions in Figures 3.1, 3.2 and 1.1, the ends are marked by a definite oscillation that corresponds to a normal breathing oscillation, and the end of the flat region can be unambiguously located to within one second. The end point was usually agreed on to within three tenths of a second.

The start time of an apnoea is usually not as easy to identify as the end time. When breathing ceases, the Graseby signal does not remain flat at its current level; it returns to a rest value, as mentioned in Section 2.3. Thus, if the infant stops breathing at the end of an expiration, where the signal is at a peak, the signal will then drift back to a mean rest value. Therefore, the movement in the signal after the peak does not represent a physical breathing movement, and the start of the pause in breathing would be the end of the expiration—the peak of the signal, as opposed to the start of the flat region. In other words, the period during which the signal decays to the rest value is a part of the pause in breathing. With high amplitude signals, the decay curve may last two to three seconds, a duration which is significant when the desired accuracy is to within one second. A 1.0 to 1.5 second decay curve has been reported by another group [Kirjavainen, et al. 1996]. An extreme example is the region *C* in Figure 3.3, where three experts agreed that *C* could represent the beginning of a period of breathing cessation. Note that it is possible that *C* is an accurate representation of breathing, in other words that the infant inhaled slowly for half a breath. (A falling signal corresponds to inhalation, described in Section 2.3.) The decay characteristic is not unique to the Graseby, and is seen in the Corometrics and airflow signals at the beginning of the apnoea in Figure 2.4 [Bruckert, et al. 1982]. The point is that there is no way of knowing whether the decay curve corresponds to breathing or no breathing, and as explained in Section 3.2.2, if in doubt the experts overestimated the duration, and therefore the start time is always taken at the beginning of the decay curve.

In summary, there is minimal difficulty in detecting flat regions, but classifying flat regions as apnoea or non-apnoea requires interpreting subtle signal shape variations. Small deviations or oscillations could mean the difference between a pause in breathing and a slight breathing movement. Determining the start and end times also requires detailed interpretation, a fact reflected in the 37% disagreement rate regarding the durations of apnoeas. Some recordings had more noise or sections of poor quality signal than others, and hence there was a higher level of disagreement between the experts with respect to the apnoeas in those recordings. Overall, it can be concluded that expert knowledge is required to accurately detect apnoeas.

3.4 Discussion and Conclusions

A methodology for human expert apnoea detection was developed, and three experts detected the apnoeas within ten recordings. The three human experts had considerable experience with reporting the results of polysomnography and were familiar with Graseby signals. They discussed in detail the criteria for the detection of apnoea, and developed a common approach to viewing the signal and measuring apnoea length. Any disagreements were discussed. Despite these strategies aimed at improving the consistency of detection, there was a 10% average and a 24% maximum

discrepancy between the three experts on which events were apnoeas. However, this is in line with Burgess who suggests a 90% agreement as a suitable level for experts within a group. Whyte found a mean difference of 11% [Whyte et al. 1992], and the 10% disagreement also compares favorably with other groups that recorded disagreement in the order of 20% [Parmalee, et al. 1972, Burgess 1990, Kahn, et al. 1992].

There are several causes of disagreement between the experts. Apnoeas are measured to the nearest second, and some of borderline length are included by some experts but not by others; these differences are based on the subtlety of deciding the exact start and end points. About half of the 10% discrepancy was due to disagreement about five second apnoeas. There was also a number of events that appeared as flat regions of signal, but did not necessarily represent apnoeas. Jeffrey et al. [1981] found events that appeared to be apnoeas exceeding 15 seconds in duration, but that were upon further analysis were found to be false detections. Such signal regions have a variety of possible causes, ranging from obvious physiological events, as in Figures 3.1 and 3.2, to signal changes with no clear causes such as a sudden flattening, as in Figure 2.13. Small variations in a flat region may represent breathing movements, indicating that breathing had not ceased completely, and although human experts sometimes recognised these variations as breathing movements, the boundary between apnoea and non-apnoea was ill-defined. In general, it appears that both physiological factors and signal characteristics led to differences in interpretation between experts.

The durations of apnoeas were disagreed upon for 37% of events. One reason the duration agreement is lower than the detection agreement could be because detection is a binary result, pause or not pause, whereas duration is based on start and end times with many sample points that could be interpreted as the correct times. Another reason could be that the experts do not detect the start and end points as part of their routine analysis, and they are not familiar with interpreting the signal in this way. A third reason could be that breathing does not start and stop in a switch-like fashion, but instead changes continuously, and therefore the exact boundary between breathing and no breathing is unclear. It is likely that all three reasons contribute to the high level of disagreement. However, the accuracy of the duration measurement is less important than the detection of apnoeas, and the level of disagreement is not considered a problem.

The breathing signals and apnoeas detected by the three experts can be used to constitute a reference set which represents combined expert opinion, such as set B_1 as explained in Section 3.2.3. The measurement of uncertainty between the experts is associated with each reference set, and therefore, when a reference set is used as a performance goal for an apnoea detection algorithm, the uncertainty can be used as an indication of what is an acceptable error for the detection algorithm. From the beginning, it is important to note that a detection algorithm based on such a performance goal cannot achieve total accuracy relative to all experts, and is therefore unlikely to achieve total accuracy relative to the combined opinion of the experts. However, the aim is to achieve the highest possible accuracy relative to the combined opinion of all three experts.

Having considered the experts' interpretations, the experts' method of apnoea detection can be distinguished as having two steps:

1. Detecting flat regions;

2. Interpreting and classifying flat regions as apnoea or non-apnoea.

The first step was relatively straight forward, as it was simply a matter of distinguishing flat regions of signal from continuously oscillating periods. However, the second step appeared to be more complex, and required detailed study and interpretation of the shape of the flat region, including start and end times. It is this second step that led to difficulties, and that does not appear to have been addressed in previous literature. These steps are considered in more detail in Chapter 6, where they are used as the basis of an apnoea detection algorithm.

In conclusion, apnoea detection by human expert is not totally accurate, and as it is the gold standard for apnoea detection, no apnoea signal definitions or detection algorithms can be totally accurate. However, to develop definitions and detection algorithms, a standard is needed against which to measure accuracy and performance, and therefore human expert opinion as it stands is used as the reference standard.

Chapter 4

Approaches to Breathing Signal Analysis

In this chapter, methods of analysing breathing signals are investigated with respect to apnoea detection. A range of methods are examined, including a review of existing algorithms for breathing signal analysis, and some methods are applied to the Graseby signal. Performance measures, as required in the evaluation of detection methods, are also considered.

4.1 Objectives

This chapter has two main purposes. Firstly, approaches to breathing signal analysis with respect to apnoea detection are considered. Secondly, the applicability of these approaches to the Graseby signal is discussed. This chapter presents a detailed background of methods that have been used to analyse breathing signals, and discusses whether or not these methods can be usefully applied to the Graseby signal.

In order to evaluate apnoea detection algorithms, measures of performance are required. A performance measure allows a method to be objectively evaluated and different algorithms to be compared. Any measure is relative to a reference set of signals and apnoeas, and of the available reference sets, B_1 is the obvious choice as it represents a verified set of events based on the opinions of several experts. B_1 is a set of 619 apnoeas found by three experts, and the associated breathing signals from ten BabyLog recordings—see Section 3.2.3. Thus, throughout this chapter, when performance measures are calculated, B_1 is used as the reference standard.

Most existing methods of analysing breathing signals are not specific to apnoea detection—as explained in Section 1.3, there are few such methods published. Although there appear to be very few apnoea scoring algorithms described in detail, the methods described here give a background of what analysis techniques have been used to analyse breathing.

There may also be signal analysis techniques that have not yet been applied to analysing breathing signals which could lead to meeting the objectives of accurately defining and reliably detecting apnoeas. This chapter does not aim to be a comprehensive evaluation of all possible analysis methods, but some approaches are adapted for the Graseby signal, and some tests performed.

Summarising, the objectives of this chapter are as follows:

1. Develop performance measures for evaluating and comparing detection algorithms;
2. Investigate methods of analysing breathing signals;
3. Apply promising methods to the Graseby signal.

4.2 *Quantifying Performance*

Quantitative measures are required to evaluate the relative performance of apnoea detection algorithms. In evaluating an apnoea detection algorithm, the following questions must be answered: How many apnoeas does the algorithm miss? How many false events does it detect? How reliable is it? The answers to these questions are critical to some applications. For monitoring, an algorithm must miss very few apnoeas and reliably detect *all* longer (>15 seconds) apnoeas. On the other hand, for apnoea scoring, a 100% detection rate is not as important as having a manageable number of false detections.

Given two algorithms, a quantitative measure is required to determine which one has better performance. Once a measure of performance has been calculated, confidence intervals can be used to give the accuracy of the measure. This section describes how performances are measured; the situation considered is using an apnoea detection algorithm to analyse a breathing signal recording, and then comparing the events detected by the algorithm to apnoeas in a reference set. These methods are used throughout the remainder of this thesis.

4.2.1 Matching Detected Events with Reference Apnoeas

Before any performance measures can be calculated, a method is needed to determine whether an event detected by an algorithm is indeed an apnoea. Therefore, any event detected by an algorithm is compared with a reference set of apnoeas. A method of deciding whether an event detected by an algorithm corresponds to an apnoea in a reference set is described: in essence, if an event occurs at the same time as an apnoea, a match is made.

The criterion for deciding when an event matches an apnoea is that there is some overlap, which is determined from the start and end times of the event and the apnoea. The start and end times of a detected event are denoted respectively as t_s and t_e, and the start and end times of a reference apnoea are denoted respectively as t_{start} and t_{end}. Thus, an overlap and therefore a match occurs if $t_s < t_{end}$ and $t_e > t_{start}$. This criterion is illustrated in Figure 4.1.

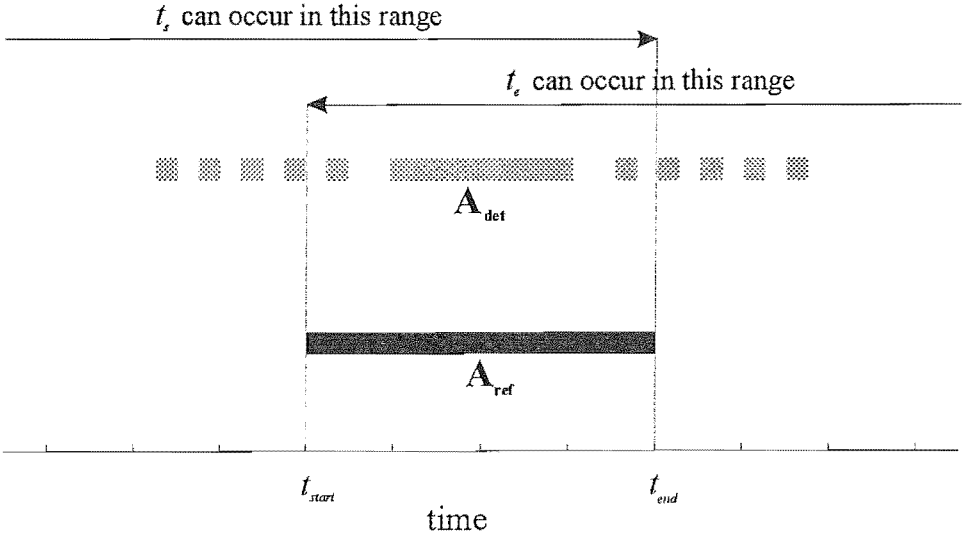

Figure 4.1 Criterion for matching detected apnoea A_{det} with reference apnoea A_{ref}. The ranges of the A_{det} start and end times t_s and t_e are shown relative to the A_{ref} start and end times t_{start} and t_{end}. For A_{det} to match A_{ref}, both $t_s < t_{end}$ and $t_e > t_{start}$ must be true. Thus, A_{det} must overlap some but not necessarily all of A_{ref}.

Using such a criterion, a detected event does not need to occur at exactly the same time as an apnoea in order to be matched to that apnoea. In fact, in the extreme case, the detected event and apnoea need only overlap by a single sample length in order to be matched. The reason that two events are matched even when they only just overlap goes back to the philosophy of apnoea detection, which is to detect and present apnoeas for evaluation by human experts. When experts are presented with detected events, they also view the signal either side of the detected event, and therefore they can recognise an apnoea even if it occurs beside the detected event. Therefore, events do not need to match apnoeas directly, and the criterion of any overlap can be used. This criterion also has the advantage of being simple and objective—there is no requirement to develop percentage overlaps or other complex measures that quantify a proportion of overlap that must be present before an event is matched with an apnoea.

Another aspect of the criterion for matching is that any one event can only be matched to one apnoea: if an event spans two apnoeas then it is matched to the first apnoea. The second apnoea is not considered detected unless overlapped by another event. If two events are matched to one apnoea, the start times of both events are compared: if they are the same, then one is eliminated and the other is matched to the apnoea. If the start times are different, the first is matched and the second is classified as a false detection.

4.2.2 False Negatives and False Positives

In quantifying the performance of an apnoea detection algorithm, the essential information is the number of apnoeas that are missed and the number of false detections, relative to some reference standard. These respective numbers are the false negatives and the false positives, and are usually described as rates, measured in percentages. The false negative rate, f_n, is the percentage of missed apnoeas relative to all reference apnoeas. The false positive rate, f_p, is the percentage of false detections relative to all events detected.

The false negative and false positive rates are with respect to a particular reference set. An algorithm that is totally accurate relative to a reference set defined by one expert is likely to have some false positives and false negatives relative to a reference set defined by any other expert. An aim of this research is to emulate expert detection *generally*, that is relative to any expert. It is more important to detect all apnoeas found by any one expert than to only detect those apnoeas agreed upon by all experts. Therefore, a low f_n is critical, and a low f_p is useful; in other words, sensitivity is more important than specificity (see Section 3.2.2).

As mentioned in Section 3.2.3, the reference set B_1 has been reviewed by several experts several times, and is the most reliable reference set of BabyLog apnoea and non-apnoea signals in terms of representing combined expert opinion. Although the false negative and false positive rates of an algorithm relative to other available reference sets may be useful, B_1 is the most general set, and the most important rates in relation to the objectives described in Section 1.4 are relative to B_1.

4.2.3 Performance Measure

The false positives and false negatives are measures of the performance of an algorithm, but they are not a single measure—how can two algorithms with different f_n and f_p be compared? A single

measure is developed that is a quantitative measure of the performance of an apnoea detection algorithm.

As mentioned in Section 4.2.2, there is a trade-off between reducing missed apnoeas and increasing false detections. To evaluate what levels of missed apnoeas and false detections are optimum, a performance measure in the form of a penalty function P is used, with lower P corresponding to better performance [Beveridge and Schechter 1970]. P is a function of f_n and f_p, and is based on the opinions of human experts regarding the relative merit of systems with different f_n and f_p.

The human experts' ratings of the performances of algorithms vary significantly depending on the values of f_n and f_p. There was a strong distinction between algorithms that were considered useful, and algorithms that were of no value. There was no exact cutoff between algorithms that were considered useful and those that were not, but in general, algorithms with low f_n and f_p were considered more useful to clinicians than those with high f_n and f_p. To accurately represent the relative merits of various performances, the penalty function P needs to represent the difference between high f_n and f_p, and low f_n and f_p, and P also needs to represent the differences between performances with similar f_n and f_p. Therefore, a function P was developed that consists of two functions: a first order linear function of f_n and f_p, and an exponential function of f_n and f_p:

$$P(f_n, f_p) = s_1 f_n + s_2 f_p + \exp(d_1 f_n) + \exp(d_2 f_p) \qquad (4.1)$$

The parameters s_1, s_2, d_1 and d_2 are tuning parameters. For high values of f_n and f_p, the performance is poor and the value of P is much greater than the value of P for low f_n and f_p

The exponential function increases rapidly, so P is small only for low f_n and f_p. As f_n and f_p increase, P increases rapidly. The exponential components of P represent the significant difference between high f_n and f_p, and low f_n and f_p. The linear function in P represents the differences between performances with similar f_n and f_p, because for low values of f_n and f_p, the linear component dominates the exponential component. A penalty function of the form of P has been used previously to determine the optimum settings for a system [Macey, et al. 1995].

In order to set the tuning parameters s_1, s_2, d_1 and d_2, a detection algorithm using approximately 60 different parameter settings was used to detect the reference apnoeas, and for each setting f_n and f_p were calculated. The different performances were studied by three experts, who then ranked them in order from best performance to worst performance. The parameters s_1, s_2, d_1 and d_2 in (4.1) were adjusted so that the 60 results ranked by P values matched the experts' ranking. Note that because detecting apnoeas is more important than reducing false detections, false negatives are penalised more heavily than false positives, and so $s_1 > s_2$ and $d_1 > d_2$. The results are shown in Table 4.1. As expected, s_1 and d_1 are greater than s_2 and d_2 respectively, reflecting the greater importance of false negatives compared to false positives.

Parameter	s_1	s_2	d_1	d_2
Optimum value	3.0	0.25	1.0	0.025

Table 4.1 Penalty function optimum parameter values.

The penalty value P as described in (4.1) allows the relative merit of different performances to be measured, with the measure relating to human experts' opinion of what is good or poor

performance. Describing P as the sum of two penalty functions S_P and D_P allows P to accurately reflect experts' opinions.

4.2.4 Confidence Intervals

Confidence intervals can be quoted as a measure of the accuracy of performance measures such as f_n and f_p. This section describes how confidence intervals are calculated. The case considered is measuring apnoea detection performance relative to a reference set of apnoeas.

A sample probability may be calculated from the number of apnoeas X detected out of all N detections. As there are two possible outcomes for each trial (an apnoea detected or not), there is an associated binomial distribution [McClave and Benson 1991]. If there is a sufficiently large number of samples, the distribution can be approximated by a normal distribution [McClave and Benson 1991]. The sample probability $\bar{p}$ is then a maximum likelihood estimator of the population probability p for large populations, and is defined:

$$\bar{p} = \frac{X}{N}. \tag{4.2}$$

Generally, this normal approximation to a binomial distribution is appropriate if the following is true:

$$0.0 < \bar{p} \pm 3\sigma < 1.0, \tag{4.3}$$

where σ is the standard deviation of the sample probability [McClave and Benson 1991]. If the assumption of the normal distribution holds, the sample probability has confidence intervals approximated by:

$$\bar{p} \pm z s_{\bar{p}}, \tag{4.4}$$

where $s_{\bar{p}}$ is the estimator of σ, and is given by:

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{N-1}}, \tag{4.5}$$

and $z(\alpha)$ is the associated standard normal value corresponding to a probability α. The value of z is determined using look-up tables [Neter et al. 1978].

If the normal assumption is not appropriate (i.e. equation (4.3) does not hold), then the exact method should be used [McClave and Benson 1991]. Such a case is likely when $\bar{p}$ is close to 0.0 or 1.0, which for the above example would be when X is close to N. With the exact method, the confidence intervals are calculated from the binomial distribution based on the observations. The binomial distribution is given by:

$$\binom{N}{X} p^X (1-p)^{N-X}, \tag{4.6}$$

where

$$\binom{N}{X} = \frac{N!}{X!(N-X)!}. \tag{4.7}$$

Hence, given X and N, a distribution curve can be calculated from (4.6) using a large number of values of p (typically 10,000), and the confidence intervals are calculated from that curve.

Confidence intervals are quoted in one of two ways, depending on how they are calculated: If the confidence interval is quoted with a plus or minus sign (for example, $\overline{p} \pm a$), then the normal distribution was assumed. If the confidence interval is quoted as a range (for example, $\overline{p}$ [a, b]), then the normal distribution was not assumed, and the exact method was used. The two possible methods of calculating confidence intervals are illustrated above using the number of apnoeas X detected out of all N detections, but the method generalises to any result with a binomial distribution (two possible outcomes).

4.3 Existing Methods of Analysing Breathing

Although there are few well-described apnoea scoring systems, several methods used to analyse infant respiration signals have been reported in some detail. The purpose of the analyses range from monitoring to measuring breathing rates, and sometimes include apnoea detection. Each research group tends to have specific functions for which their system is designed. These methods of analysing breathing signals are reviewed and evaluated in terms of the objectives in Section 4.1. Some methods that have been specifically applied to apnoea detection are also introduced, with evaluations of these methods presented in later sections.

An algorithm designed for apnoea scoring has been developed by Peirano et al [1988], and was used as a computer tool to assist clinicians. The algorithm uses a semi-automatic adaptive threshold in order to detect apnoeas. This method involves calculating the mean signal amplitude for each of 20 consecutive one second intervals and also the mean amplitude over the entire 20 seconds. If more than three consecutive one second intervals are less than 25% of the 20 second interval mean, then the three intervals are tagged as a possible apnoea. This possible apnoea is then left for an expert to classify. Although not specified, the breathing signal appears to be chest volume measured using impedance pneumography. The system does not appear to have been tested without expert input, and it is implied that the system detects many false positive events [Peirano, et al. 1988].

As well as apnoeas, breathing rate is another variable that is calculated from breathing signals. Breathing rate can be calculated from breath lengths. Breath length can be measured using a method of detecting zero-crossings, as a zero crossing will occur at the rising and falling of each breath oscillation. In its basic form, the zero-crossing method is susceptible to noise. Therefore, Ning and Bronzino [1989] have used a modified zero-crossing method to estimate the respiratory rate, along with an "energy index," which they define as a measure of the power of the breathing. An apnoea is detected when the energy index is low. Ning and Bronzino reported that the system achieved 100% detection of 33 apnoea events from almost one hour of breathing recorded from an adult. However, they failed to specify the type of signal used and the nature of the events. As adults do not normally have apnoeas, let alone 33 in one hour, it is likely that the apnoeas were conscious breath-holding spells. Such test data means that the system was only tested to a limited extent, and the performance may not generalise to all signals. As explained in Sections 2.3.3 and 3.1, there are many events during infants' sleep where breathing almost stops, and these events are often a source of false detections. Another common occurrence in infants' recordings is low amplitude breathing, and if Ning and Bronzino tested the algorithm on awake

adults then no low breathing amplitude would be present. The authors themselves state: *It is our opinion that an optimal detection and classification scheme can only be developed when agreements upon performance measures and benchmark signals are standardized* [Ning and Bronzino 1989].

Another method of calculating breathing rate has been proposed by Moyles et al. [1989], who present a statistical approach to segment an airflow breathing signal into breaths. The breaths are detected by segmenting the signal into inspiration and expiration regions, and measuring the duration between transitions. The segmentation is performed by classifying a region of signal by the trend in values, not including the values about the middle region of the window. If, on average, the signal is rising then the trend is up (inspiration); if, on average, the signal is falling then the trend is down (expiration). Using trends prevents minor peaks or troughs being mistaken for transitions between expiration and inspiration. However, Moyles et al. [1989] presented no test results and no discussion in terms of their approach to apnoea detection was given.

Wilks and English [1994] applied Moyles' algorithm to a breathing signal produced using an abdominal pressure capsule (the same type of signal as the Graseby), but found that the results were inconsistent due to greater noise (see page 55). Hence, they developed an improved version of Moyles' algorithm which detects transition points more consistently. The improved algorithm proposed by Wilks and English ensures that each breath lasts at least a minimum duration, and that a transition is not immediately followed by another transition. Having found segments of inspiration and expiration (one breath), the amplitude of the signal at the adjacent inspiration and expiration segments (adjacent breaths) is compared with the average amplitude over the previous ten breaths. The two segments (one breath) are classified as a "high," "in range" or "low" breath. The breathing is classified based on the sequence of high, in range, or low breaths. The algorithm also classifies a region as "ineffective breathing" wherever the signal has an amplitude range of less than one sixth of the mean range of the previous ten breaths. The system is later combined with a neural network to classify breathing as "effective" and "ineffective," as opposed to apnoea or non-apnoea [Wilks and English 1995]; "ineffective" breathing appears to be related to, but different from, a cessation of breathing. Given such a criterion, the non-apnoea events as shown in Figures 2.8, 3.1 and 3.2 would be likely to be classified as ineffective breathing, and therefore ineffective breathing as defined by Wilks and English does not correspond to apnoea. However, as Wilks and English designed the system for real-time monitoring, they may not have placed any importance on the fact that these non-apnoea events were detected. Overall, the system appears to have high sensitivity at the expense of detecting events that are not true pauses in breathing.

Methods based on frequency techniques have been used to analyse acoustic breathing signals. Tracheal sound has been suggested as a less invasive measure of airflow compared to usual techniques [Beckerman et al. 1982], and there have been various studies into the use of breath sound as a respiration signal [Beckerman, et al. 1982, Werthammer et al. 1983], including the detection of obstructive apnoeas [East and East 1985]. A spectral analysis method has been used to detect central apnoeas from a breath sound signal [Ajmani et al. 1996]. The essence of the method developed by Ajmani et al. is that the energy of the acoustic signal during breathing is high, and the energy during a pause is low; hence a low energy spectrum should correspond to an apnoea. The signal is sampled at 2.2kHz and then band-pass filtered between 150 and 600Hz. The

spectral density energy is calculated over epochs of four seconds duration, and if below a threshold, an apnoea is detected. The system has been tested on three recordings of adult breathing, including simulated apnoeas. The results showed few false positives (0% in two out of three studies), but many false negatives (20% or more in all three studies). This high false negative rate reflects difficulties in analysing acoustic breathing signals, and the lack of guidelines and standards in this area [Mussell 1992].

There are some methods that use multiple signals to detect apnoeas, either several breathing signals or a variety of signals such as breathing, blood oxygen saturation and heart rate. Marcotte et al. [1996] present a method that uses features from each signal to classify one event as apnoea or non-apnoea. Another method analyses the frequency characteristics of heart rate as they change over time, a similar method to the spectrogram (see Section 4.5) [Afonso et al. 1994]. The spectral information obtained is used to improve the performance of a breathing signal-based apnoea detection algorithm, but not as a stand-alone analysis. Combining signals is clinically useful, and may be of use in the future in order to develop a more accurate apnoea detection system. Rakowski et al. [1986] use an airflow signal initially to detect possible apnoeas, and then use several thoracic and abdominal signals to classify the apnoea. However, the methods of detecting features in the individual signals are little different from methods that analyse only one signal. Furthermore, the methods that use multiple signals do not suggest possible improvements to apnoea definitions or detection algorithms based on a single breathing signal.

Revow et al. [1986] has described a method to calculate breathing rate using a peak-to-peak detection method, a method that had been briefly described some years prior [Mason, et al. 1974]. Revow et al.'s method performed well in terms of detecting peaks when tested on ten hours of chest volume breathing data [Revow, et al. 1986]. Their method relies on the fact that the absolute amplitude of the breathing signal being analysed corresponds directly to ventilation. This is not a characteristic of the Graseby signal, as the Graseby has variable amplitude that changes throughout a recording (see Section 2.3.3). Mason's et al.'s [1974] method was applied to both airflow and impedence signals, and appears to have been successful in detecting peaks in variable data. Essentially, peaks were verified relative to previous troughs by ensuring that they were separated by a significant amplitude range, where the amplitude range was relative to the signal. Although this algorithm was in clinical use, no further results were presented. Another peak-to-peak detection method of calculating breathing rate is presented by Wilson et al. [1982]. This method uses a series of peak detection algorithms, and includes various checks to ensure that peaks detected are true peaks and not due to heart beat artifact, body movement, or other insignificant signal movements [Wilson, et al. 1982]. In fact, Wilson's method is similar to another peak detection algorithm used for calculating breathing rate [Tuffnell 1993]. However, with all of these methods, no evaluation of apnoea detection performance was presented. Because these peak-to-peak methods have been used for some applications, and there are several published algorithms that use this approach, they are further considered in Section 4.4.

A method used to analyse a chest breathing signal for apnoea monitoring has been developed based on a combination of peak-to-peak detection and features of the shape of the breathing signal [Sahakian and Tompkins 1982]. Sahakian and Tompkins' algorithm firstly detects breaths using peak-to-peak detection, which is similar to other methods [Mason, et al. 1974]. Having

detected a breath, the slope of the signal from trough to peak or vice-versa is then measured using five-point parabolic fits. If the slope exceeds a threshold, then motion artifact is suspected and that breath is discounted. No breaths are detected until the slope returns to normal limits for a period. The slope threshold is defined as twice the average of the slope during the last four accepted breaths. An apnoea is detected if fewer than eight breaths are detected within a period of 15 seconds. The performance of the method reflects the monitoring aspect of the algorithm, as apnoeas are not located precisely nor are their durations measured. No results are given [Sahakian and Tompkins 1982].

An apnoea detection algorithm has been developed to analyse an expired CO_2 (airflow) signal for monitoring purposes [Laxminarayan, et al. 1982]. This method is based on the premise that time domain methods are ineffective, and frequency domain methods show more promise. The algorithm applies Walsh transformations, which are in essence a combination of zero-crossing detection and frequency analysis [Beauchamp 1975]. Walsh transformations give an estimate of the spectrum, and are computationally more efficient than the FFT. The Walsh transformation analysis detects zero-crossings about a mean, and from the durations of the intervals between zero-crossings, an estimate of frequency is calculated. If the Walsh power spectrum is low, then an apnoea is said to have occurred. The purpose of this algorithm was to identify apnoeas longer than 15 seconds, and no attempt was made to detect shorter apnoeas (referred to by the authors as "respiratory pauses") or to measure the duration of the identified events [Laxminarayan, et al. 1982]. The method is designed for real-time monitoring, and performed well on five test recordings, with 1-2% false positives and 2-3% false negatives. If applied to a Graseby signal, the system may have difficulty identifying low amplitude breathing. The greater noise in the Graseby signal compared to an airflow signal could lead to spurious zero crossings being detected, and hence cause errors in the Walsh power spectrum, as when Wilks attempted to apply Moyles' zero-crossing algorithm to an abdominal signal [Moyles, et al. 1989, Wilks and English 1994]. As with other methods, flat regions that are non-apnoea events (Figures 3.2 and 3.3) are likely to be detected as apnoeas. Although Laxminarayan's method is described with a mathematical rigour that has been lacking in descriptions of apnoea scoring methods, this method is likely to have the same problems as other techniques when applied to the Graseby signal for apnoea scoring.

Rakowski et al. [1986] present a method based on an airflow signal, where an apnoea is defined as occurring where a signal is "straight" for at least ten seconds, or whatever minimum duration is set. They then had the problem of defining straight, a signal property referred to as flat by other groups [Laxminarayan, et al. 1983, MacFadyen, et al. 1988]. As with these other groups, they defined a straight as being below a percentage of the amplitude range of breathing. In their case, this threshold was set at the beginning of the recording by an user, and was usually in the order of 20%. The method was applied to airflow signals only, and was reported to produce "good results," but no test results were given [Rakowski, et al. 1986].

Bruckert et al. [1982] present a similar method which defines flat regions as regions of signal where the derivative of the signal is within a range about zero. The method calculates the start of the pause in breathing as the peak preceding the flat region, a definition that relates to that of the human experts described in Section 3.2.1. The flat region is set to be at least 0.5 seconds in

duration, even though three seconds was the shortest duration of breathing pauses being detected. The overall results were a false negative rate of 24% and a false positive rate of 17% [Bruckert, et al. 1982]. The false positive rate was reduced to less than 4% by representing the false positive events to the expert, who then reclassified the majority of these as apnoeas.

In summary, the existing methods can be separated into three categories:

1. There are several methods that segment respiration into breaths [Wilson, et al. 1982, Revow, et al. 1986, Moyles, et al. 1989], of which some are also used to detect apnoeas [Ning and Bronzino 1989, Wilks and English 1994, Wilks and English 1995]. The primary aim of the majority of the methods reviewed is either breath rate calculation or apnoea detection for monitoring. The existing methods are not designed to specifically discriminate between true apnoeas and the similar non-apnoea events. Of those methods reviewed, the effective apnoea scoring systems include the statistical method developed by Peirano which analyses a chest impedance signal, but the system was presented only as a tool to assist clinicians, and was not evaluated as an automatic apnoea detection algorithm [Peirano, et al. 1988]. Other systems that appear to have been used effectively for apnoea detection include the methods based on peak-to-peak measurement of breath lengths, although few results were given [Mason, et al. 1974, Sahakian and Tompkins 1982, Wilson, et al. 1982, Revow, et al. 1986]. Such a method is evaluated in Section 4.4.
2. Methods based on the frequency characteristics of breathing signals have been used, although to a limited extent [Beckerman, et al. 1982, Werthammer, et al. 1983, Ajmani, et al. 1996]. Frequency techniques of apnoea detection are explored in Section 4.5.
3. Methods that analyse the signal flatness tend to be specific to apnoea detection, with some results given [Bruckert, et al. 1982, Rakowski, et al. 1986]. Another point about these types of analyses is that they are similar to the definitions that human experts used to describe apnoea signals (see Section 3.2.1) [Laxminarayan, et al. 1983, Butcher-Puech, et al. 1985, MacFadyen, et al. 1988, Kahn, et al. 1992]. This type of analysis is investigated in detail in Chapter 5.

There has also been an initial investigation of neural network analysis, but the results were inconclusive [Sturman 1991]. A neural network approach to apnoea detection is therefore considered in Section 4.6. Overall, there are few systems specifically designed for accurate apnoea scoring.

4.4 Evaluation of a Peak-to-Peak Apnoea Detection Algorithm

Peak-to-peak apnoea detection is one method of computerised apnoea detection that has been in clinical used [Mason, et al. 1974, Hoppenbrouwers, et al. 1977]. Mason et al.'s [1974] method has been applied to an airflow signal, but no detailed performance figures were published, although it has been reported to perform satisfactorily [Hoppenbrouwers, et al. 1977]. This section evaluates a peak-to-peak apnoea detection method as applied to the Graseby signal, including measures of performance.

Peak-to-peak detection algorithms have been used to detect breaths, with the peak of a signal corresponding to the end of inspiration or expiration [Marshall 1986]. Thus, the time between two peaks is the length of the breath, and the rate can be calculated by taking the inverse of the breath duration. Typically, the median breath rate is used, which is defined as the median rate over a window [Harper, et al. 1987]. As peak detection algorithms have been used for breath rate measurement, it was a logical extension to apply them to apnoea detection.

An apnoea can be considered an extended breath, and therefore apnoea detection is performed simply by detecting all breath lengths greater than a minimum duration. This definition of apnoea has been used clinically by human experts detecting apnoea from breathing signals [Kahn, et al. 1988]. It has advantages in that the peaks are usually clearly defined and can be located precisely. There are certain problems with detecting peaks in biological waveforms, but these have been well-described elsewhere [Mason, et al. 1974, Marshall 1986, Tuffnell 1993]. In essence, breathing signals have variable amplitude range, and they have spurious peaks due to signal noise and physiological noise, such as cardiac oscillations. Furthermore, shallow breathing can produce signals of very low amplitude [Tappin, et al. 1997].

Another problem with peak-to-peak apnoea detection is that the duration of the time between peaks is not a measure of the pause in breathing. The peak is the end of expiration in the case of the Graseby signal, and there is therefore possibly going to be some inspiration at the start of the apnoea time, and some expiration at the end of the apnoea time. It has been found that peak-to-peak apnoea detection tends to overestimate the duration of a pause in breathing by approximately one second [Hoppenbrouwers et al. 1980a, Hunt, et al. 1988].

A peak-to-peak algorithm was used to analyse the Graseby breathing signals in reference set B_1 [Tuffnell 1993]. Tuffnell's algorithm is similar to Mason et al.'s [1974] in that it takes into account the potential problems with breathing signals mentioned above—spurious peaks, changing amplitude and cardiac oscillations. It has been successfully used to measure median breathing rate over many studies, and has consistently produced reliable results in terms of breath rate calculation [Tappin, et al. 1996a].

The breathing signals in B_1 were analysed using the algorithm, and every breath length (time between adjacent peaks) was recorded. All breaths greater than five seconds were classified as apnoea, and compared with the expert apnoeas. As peak-to-peak apnoea detection algorithms have been found to overestimate the length by one second [Hoppenbrouwers, et al. 1980a, Hunt, et al. 1988], this procedure was repeated for all breaths greater than 6 seconds in duration. The results are shown in Table 4.2.

Peak-to-peak	Missed Apnoeas (out of 619)	False Detections	False +ve	False -ve
5+ second breaths	0	10,379	94.4%	0.0%
6+ second breaths	17	5,861	90.7%	2.7%

Table 4.2 Peak-to-peak performance compared to reference set B_1.

The results show that the algorithm performs poorly, especially in terms of false positives. Even when only taking six second breaths, almost 6,000 false positive events were detected, making the algorithm impractical for use on the Graseby signal. This large number of false

detections is caused by the signals characteristics described in Sections 2.3.3 and 3.3, and is further evidence to support the development of specific apnoea detection algorithms.

Another factor in the reliability of the algorithm is that the entire Graseby recordings were analysed, and no data were excluded. When Mason et al's [1974] algorithm has been used, long episodes of crying have been deleted from the analysis (up to 10%) [Hoppenbrouwers, et al. 1980b]. It is possible that if some of the Graseby data was removed from the analysis, the performance of the algorithm would improve. However, the aim of this research is to develop an completely automated detection system, and hence all the data are considered in any analysis.

4.5 Fourier-Based Methods

Some pattern recognition problems are simplified when the signal within which the pattern occurs is mapped to another domain; a common example of a mapping is from the time to frequency domain. In the context of breathing signal analysis, an apnoea is a low frequency component of breathing, and therefore spectral methods can potentially be used to detect apnoeas by detecting low frequency components. Laxminarayan et al. [1982] used frequency characteristics of breathing to detect apnoeas in a monitoring situation from an airflow signal. Zero crossing techniques are closely related to frequency analyses [Ning and Bronzino 1989], and as described in Section 4.3, spectral analyses have been used to analyse acoustic breathing signals [Ajmani, et al. 1996]. Breathing signals have been directly analysed, including periodic breathing, but not specifically for the purpose of apnoea detection [Nugent, et al. 1983].

The Fourier Transform describes the mapping of a signal from the time domain to the frequency domain [Brigham 1988]. The Graseby breathing signal has been studied using Fourier transforms, with the aim of describing patterns that distinguish between apnoea and non-apnoea signals. The Fourier Transform of a sampled signal can be calculated using the Fast Fourier Transform (FFT) [Brigham 1988].

4.5.1.1 Spectrum

Frequency analyses are typically used where there is a repeating pattern of oscillations. In the case of breathing, a typical signal contains repeated oscillations corresponding to breaths. However, an apnoea is an isolated aperiodic event (unless within periodic breathing—see Section 1.1). During a night, a baby has tens of thousands of breaths but usually only 50 to 200 apnoeas, and hence any low frequency component due to apnoeas is insignificant within a spectrum of a whole night's breathing. For an apnoea to be represented by a significant peak in a spectrum, the duration of the apnoea must be a significant proportion of the window length over which the spectrum is calculated.

Most apnoea definitions include a minimum duration; five seconds was used in Chapter 3. Minimum durations used to define apnoeas are usually at least three times longer than the average breath length; in fact, a minimum duration of three times the median breath length has been used as a minimum apnoea length [Mason, et al. 1974, Tappin, et al. 1997]. The raw breathing signal recorded by the Graseby monitor is bandpass filtered between 0.67Hz and 2.3Hz, based on the assumption that infants' breathing varies between 40bpm and 140bpm (Section 2.3). However, an apnoea is a lower frequency component within a breathing signal, as the flat region

of the apnoea signal is a low frequency component relative to the surrounding breathing. If an apnoea is three times longer than surrounding breaths, then it would correspond to frequencies in the range of 0.22Hz to 0.76Hz for breathing rates between 40bpm and 140bpm. Therefore, the frequency component corresponding to an apnoea is usually one third or less than the frequency of surrounding breathing, and hence frequency techniques can potentially isolate the low frequency component of a breathing signal due to apnoeas. The frequencies of interest are close to 0Hz, with a five second apnoea corresponding to a frequency of 0.2Hz [Werthammer, et al. 1983].

Examples of breathing signals and their spectra are illustrated in Figure 4.2. Within a breathing signal, an apnoea appears as a flat region between a series of oscillations corresponding

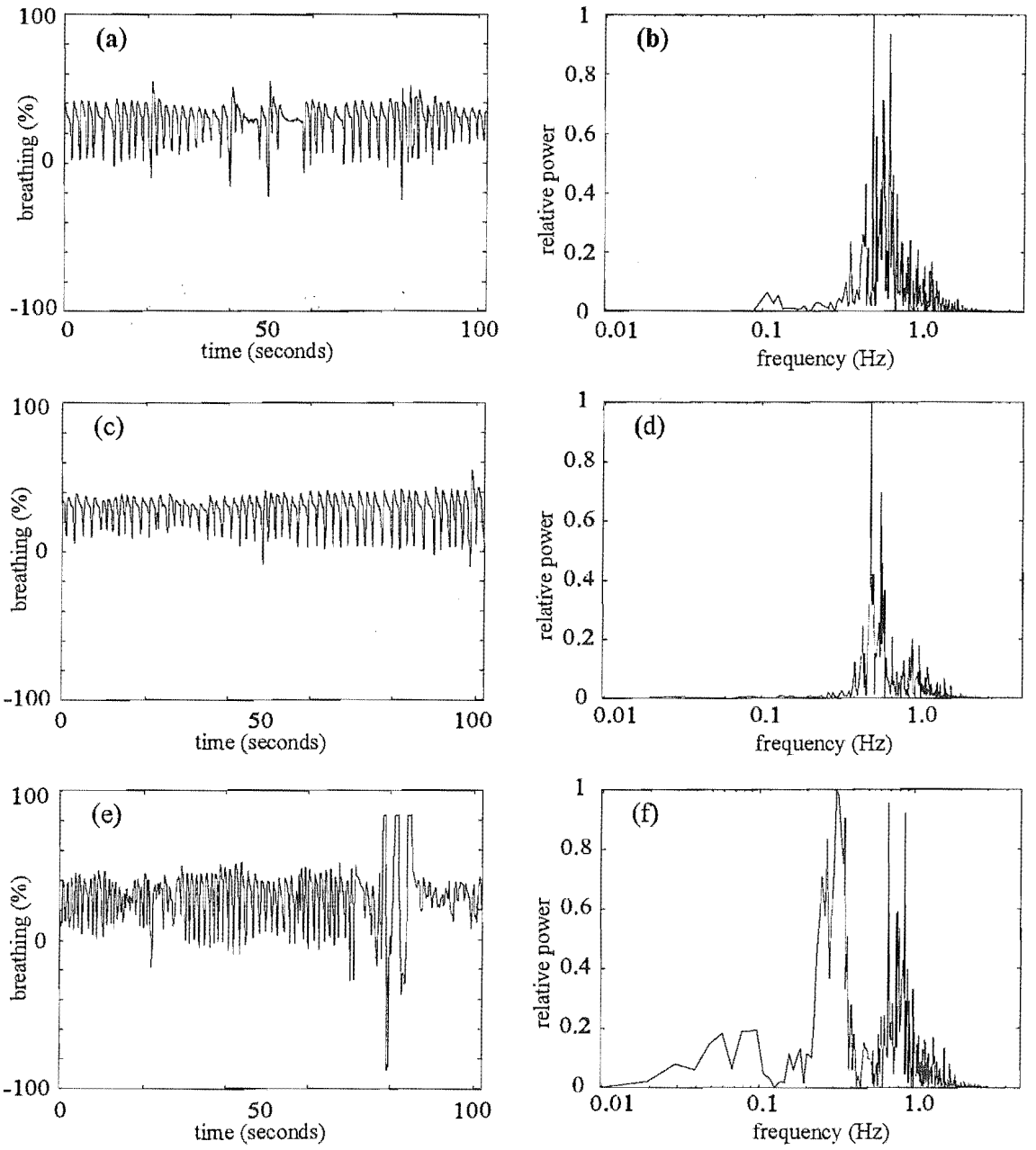

Figure 4.2 Spectra of breathing signals including apnoeas, regular breathing, and active breathing. The two apnoeas in the regular breathing signal in (a) are represented by a peak around 0.1Hz in the spectrum in (b). There is no such peak in the spectrum in (d) of the regular breathing in (c), but for the active breathing in (e), the spectrum in (f) illustrates peaks around the 0.1Hz frequency.

to breaths, and the apnoea is therefore a low frequency component. This low frequency component is seen in Figure 4.2(b) at 0.1Hz, and is distinct from the spectrum of regular breathing as in Figure 4.2(d). However, the spectrum of active breathing (see Section 2.3.3) in Figure 4.2(e) also has peaks around 0.1Hz, as seen in Figure 4.2(f). The spectra in Figure 4.2 were calculated for a window length of 102 seconds (1020 samples), and other window lengths result in similar results in terms of spurious peaks in the spectra of active breathing signals.

If defining peaks in the frequency domain that correspond to apnoeas is a simpler task than defining flat regions in the time domain that correspond to apnoeas, then Fourier-based methods could lead to accurate definitions and detection algorithms. However, based on the spectra of a variety of breathing signal types, such as those illustrated in Figure 4.2, analysing the spectra of breathing regions does not appear to offer any advantage in the context of apnoea detection.

4.5.1.2 Spectrogram

A spectrogram is a two-dimensional representation of the spectrum over time, and has been used as a tool to analyse time-domain signals [Ghitza 1994, Schroeter and Sondhi 1994]. Spectra are plotted as greyscale bands one beside the other, giving a visual representation of the spectrum changing over time. The spectrogram is calculated using overlapping windows, with each window of breathing overlapping the next such that each data sample is included in two or more windows [Schroeter and Sondhi 1994]. Any event is therefore represented in several spectra, and if there is a high degree of overlap, the spectral peaks corresponding to an event would be repeated on adjacent spectra and appear as a line. A spectrogram of breathing would be expected to show a continuous line of peaks around the normal breathing frequency, and an apnoea would appear as a line of peaks at a lower frequency. Thus, even if a peak corresponding to an apnoea was not present in a particular spectrum, it could still show up in the spectra of adjacent windows, and hence as a line of peaks in the spectrogram.

Spectrograms of breathing signals were calculated using standard parameters [Ghitza 1994, Schroeter and Sondhi 1994]. Two examples of breathing signals and their spectrograms are shown in Figure 4.3. The first example shows an apnoea during regular breathing. The spectrogram of this signal displays a line or area about 0.1Hz corresponding to the apnoea, and this area is distinct from the remainder of the spectrogram. The second example illustrates two adjacent apnoeas during active breathing. The spectrogram of this signal shows more than one line or regions that occur about the 0.1Hz frequency—these extra regions are similar to the peaks seen in the single spectrum of active breathing shown in Figure 4.2(f). In terms of detection, the apnoeas can be seen in the spectrogram but detecting apnoeas from a spectrogram is little simplified from the original problem of detecting apnoeas from a time-domain breathing signal.

4.5.1.3 Cepstra

Another Fourier-based method of representing a signal is the cepstrum. The cepstrum is calculated by mapping the signal to the frequency domain, taking the logarithm, and then mapping it back to the time domain [Oppenheim et al. 1968, Schafer and Rabiner 1970]. The cepstrum has traditionally been used in the area of speech analysis [Oppenheim, et al. 1968, Schafer and Rabiner 1970]. An advantage over the spectrum and spectrogram representations is

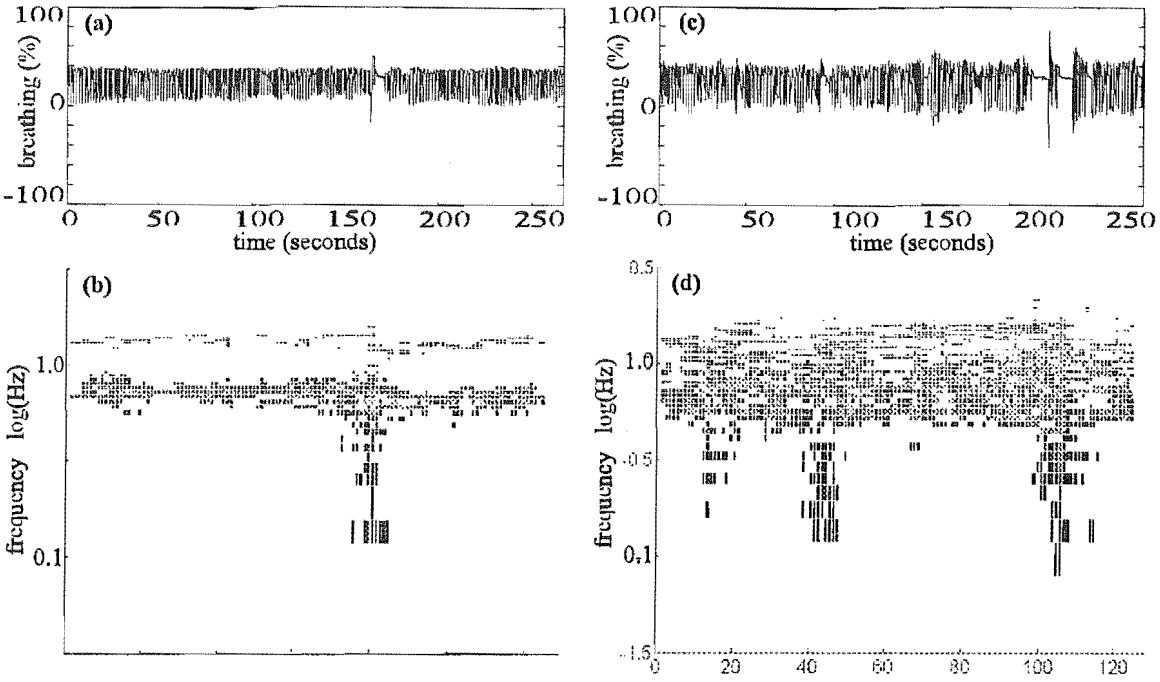

Figure 4.3 Spectrograms of breathing signals, with graphs (a) and (b) illustrating an apnoea within a period of regular breathing, and graphs (c) and (d) illustrating a period of active breathing that includes two apnoeas. Note that in (b), there is a low frequency component just above 0.1Hz which represents the apnoea in (a), but that in (d), there are low frequency components that do not correspond to the apnoeas in (c).

that the cepstrum domain is in time units, meaning that peaks representing apnoeas would be spread out from the origin, and not close to the origin as with the spectrum.

The true cepstrum, termed the complex cepstrum, involves calculating the complex log of the spectrum, which requires computationally intensive algorithms [Tribolet 1977, Sokolov and Rogers 1993]. In order to speed up processing, the log magnitude may be calculated, giving the cepstrum (as opposed to the complex cepstrum) of the signal. The cepstrum carries information only about the even features of the complex cepstrum [Noll 1967, Markel 1972]. Therefore, only the positive time cepstrum was considered, as the negative time domain of the cepstrum carries the same information as the positive time domain.

The cepstra of sequences of overlapping windows of breathing data were calculated. A Hamming window was used to smooth the data, as used by other groups [Noll 1967, Oppenheim, et al. 1968, Schafer and Rabiner 1970]. The amplitude of the breathing segment was normalised between -1.0 and 1.0, with the extreme values representing the maximum possible range from the instrument. A Fast Fourier Transform (FFT) of the data gave the real and imaginary parts of the spectrum, from which the magnitude was calculated. The log of the magnitude gave the log magnitude spectrum, which was transformed to the cepstrum via an inverse FFT.

A variety of window lengths were tested, with the window length being a trade-off between the clarity of the peaks (requiring a longer window) and the magnitude of the cepstral component due to any apnoea (requiring a shorter window length). Each window of data was filtered by a Hamming filter [Noll 1967, Oppenheim, et al. 1968, Schafer and Rabiner 1970], which smoothly reduced the amplitude of the end regions of the window to zero. A 51 second window length (510 samples) was found to give clear peaks that had a significant magnitude. Given that the middle

portion of a 51 second window would be accentuated by the Hamming filter, the 51 second window length could be considered similar to the 30 second window length that human experts used to detect apnoeas (Section 3.2.2). An example of the cepstrum of an apnoea is shown below in Figure 4.4.

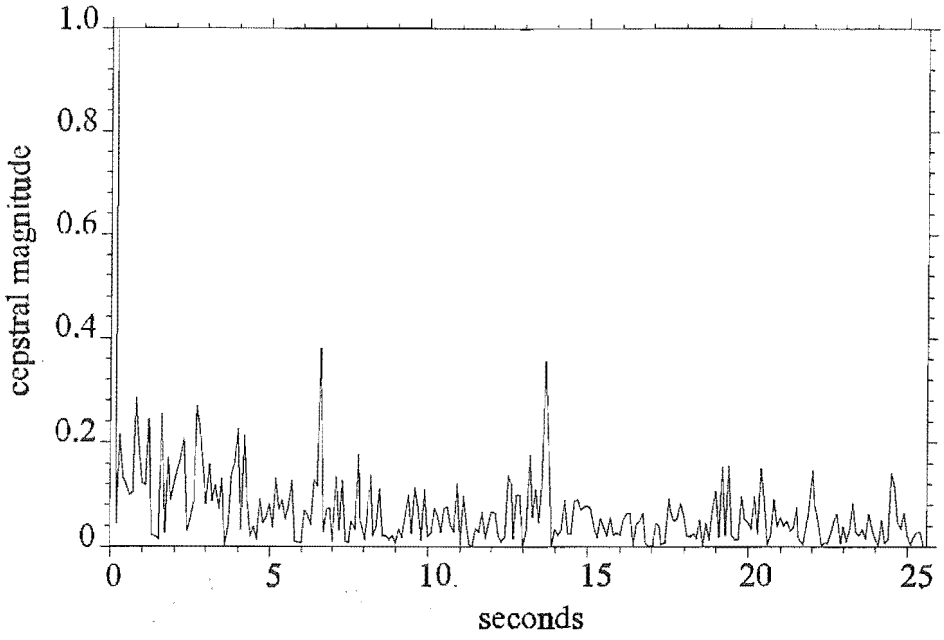

Figure 4.4 Cepstrum of a 51 second window of breathing signal that contains an apnoea. Peaks at approximately 6.5 seconds and 14 seconds are characteristic of cepstra of other apnoea signals.

Once cepstra were obtained, the aim was to find features that distinguished whether an apnoea had occurred within the period of breathing being analysed. The features of a cepstrum are peaks, as with spectra. Features of apnoea and non-apnoea within cepstra of breathing signals were found, and selected according to distinguishing features between apnoea and non-apnoea. The cepstra of ten segments of breathing with an apnoea located approximately in the centre of the segment, and of ten segments of normal breathing (non-apnoea) were calculated.

Ten cepstra of regular breathing with and without apnoeas are shown in Figure 4.5. There is no obvious single time at which peaks occur which definitively marks an apnoea or no apnoea. The graphs were divided into regions based on position (time in seconds) and amplitude. The distinguishing features of the cepstra of apnoeas and non-apnoeas for four regions are described:

	Region	Apnoea Features	Non-apnoea features
1	0.4 to 0.8 seconds	Highest peak (0.3 to 5 seconds) amplitude < 0.34	Highest peak amplitude > 0.36
2	5.0 to 10.0 seconds	Highest peak amplitude < 0.2	Highest peak amplitude > 0.2
3	10.0 to 15.0 seconds	Highest peak amplitude > 0.28	Highest peak amplitude < 0.21
4	17.5 to 25.0 seconds	Highest peak amplitude > 0.18	Highest peak amplitude < 0.14

A trial apnoea detection algorithm was implemented and tested using these features. A breathing signal was split into overlapping windows, and the cepstrum of each window was calculated and tested for the presence of the above features. If apnoea features were present, an apnoea was detected, and if non-apnoea features were present, no apnoea was detected.

The recording for patient 3 as described in Section 3.2.3 was analysed, as this patient had the greatest number of apnoeas: three human experts found a total of 145 apnoeas in this recording. The cepstral analysis detected 62 of these apnoeas (57% false negative rate) and also detected 420 other events as apnoeas (87% false detection rate). The analysis time on an IBM PC (25MHz 486) was approximately one hour for a 16 hour recording. Thus, the trial analyses performed poorly.

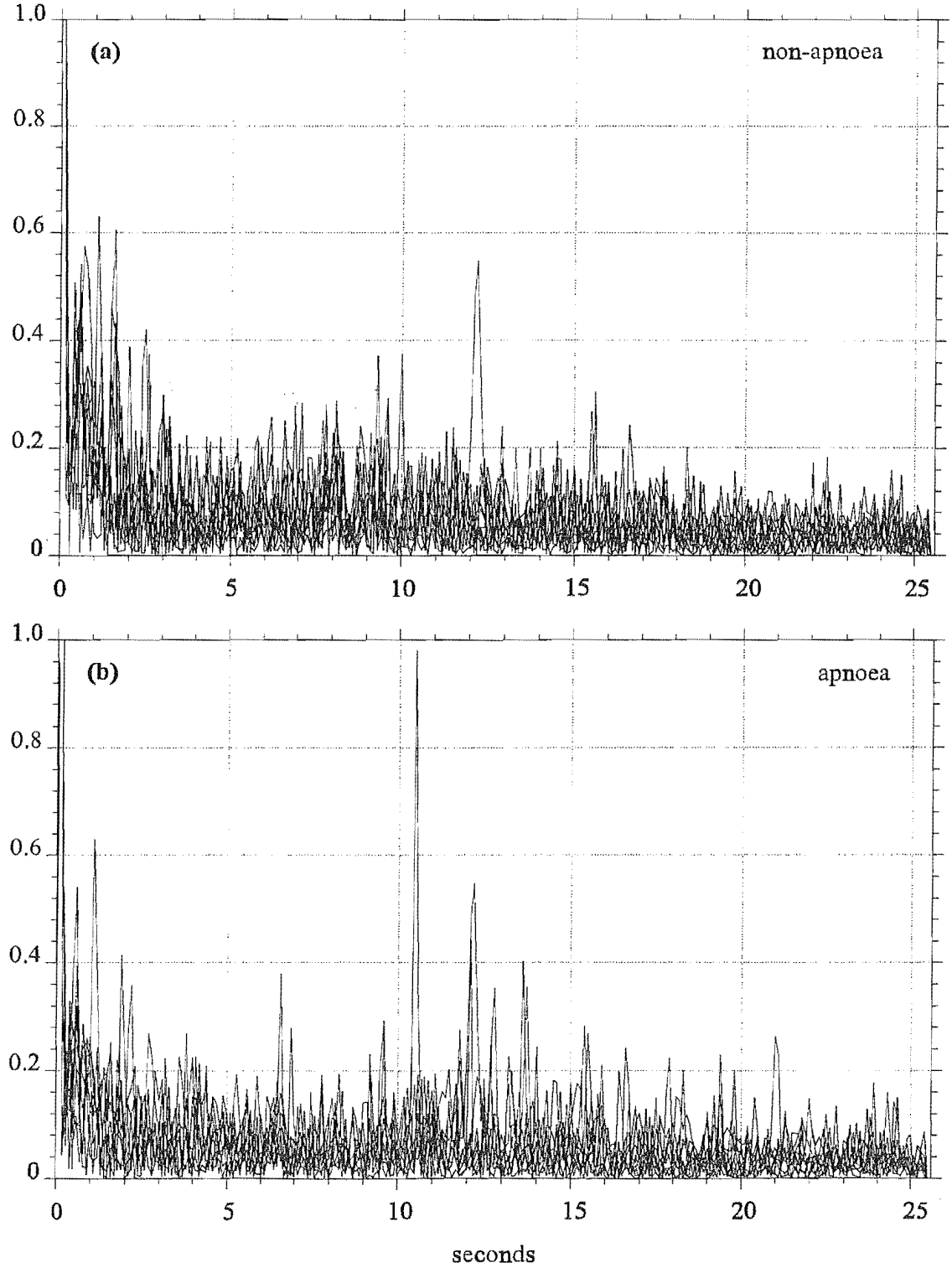

Figure 4.5 Cepstra of ten segments of regular breathing not including an apnoea (a), and ten segments of regular breathing including an apnoea (b).

The test examples were limited in number and variety, and more detailed analyses could have lead to improved performance. However, even when only considering regular breathing signals, the cepstra of the apnoea and non-apnoea signals did not show any clear distinguishing features. The cepstrum did not illustrate any features that distinguished between apnoea and non-apnoea any more so that time domain features.

Summarising, Fourier-based analyses appear to offer little in terms of improving apnoea detection. The mapping to the frequency domain does not enhance distinguishing features of apnoeas within Graseby breathing signals.

4.6 Neural Networks

Neural networks have been widely applied to pattern recognition problems, especially problems where some expert knowledge is required [Pao 1989]. Neural networks are suited to problems where the solutions are not exact, and where no exact analysis techniques available [Haykin 1996].

Neural networks are studied because they appear suited to the inherent uncertainty and lack of formal analyses within the field of apnoea detection. Neural networks have been applied in the general area of apnoea detection. Wilks and English [1995] have used a neural network to classify infants' breathing based on a set of measured parameters. A neural network has been successfully used to analyse EEG data in order to detect adult Obstructive Sleep Apnoea using measures of features of several breathing signals [Siegwart et al. 1995], but the breathing data was not analysed directly. An initial study using neural networks to analyse a Graseby breathing signal has also been performed [Sturman 1991].

This section presents research investigating the application of neural networks to apnoea detection. It includes testing both network and training parameters using simple problems, and implementing a similar network as that presented in previous work [Sturman 1991]. An introduction to neural networks is presented, and terms that are used later in this thesis are defined.

4.6.1 Introduction to Neural Networks

Neural networks are powerful, robust classifiers that are useful for problems where conventional techniques require extensive computational power, excessive storage, or complex analytical definitions [Burr 1988, Lippmann 1989, Pal and Mitra 1992].

4.6.1.1 Description of Neural Networks

An neural network is a network of simple processors that are highly interconnected. The processors are simple threshold units, summing inputs and passing these through a non-linear threshold function to produce an output. The processors are usually highly connected, with the output of one processor used as the input to other processors. Each connection has a weighting, and the weightings in a network determine the function of the network. Neural networks also have external inputs and outputs. A neural network implemented on computer or in electronic hardware is known as an artificial neural network, as opposed to a biological neural network such as the human brain. Artificial neural networks are also known as parallel distributed processors,

or connectionist models. Hereafter, the term "neural network" is used to refer to an artificial neural network.

Neural networks are based on the composition of the human brain, which consists of billions of highly interconnected neurons. A neuron is a simple biological processing unit, receiving inputs from other neurons, weighting these, and summing them to produce an output through a non-linear threshold function. Hence, the processors in the artificial neural networks are referred to as neurons, or nodes. Each node can be represented mathematically as:

$$o = f\left(\sum_{j=1}^{N} i_j w_j\right) \tag{4.8}$$

where o is the output of the node, i_j is the input from the j^{th} node, w_j is the weighting of input i_j, N is the number of inputs and f is a non-linear threshold function. The function f is referred to as the *transfer function* [Pao 1989].

The manner in which the nodes are connected defines the *architecture* of the network. A common architecture is the *layered feedforward* architecture [Pao 1989]. In a layered feedforward network, the nodes are arranged in layers, with the outputs of the nodes on one layer used as the inputs to the nodes in the next layer. The first layer is the *input layer*, which receives the external inputs to the network, and the last layer is the *output layer*, which gives the external outputs of the network. The layers between the input and output layers are called *hidden* layers. A feedforward network with hidden layers is the most common architecture type used for engineering applications [Lippmann 1989].

Neural networks can perform many functions such as pattern classification, memory recall and neurological modeling. The function of a neural depends on the architecture and the weightings (w_j in (4.8)). To perform a desired task, the values of the weights are adjusted so that the network gives a desired output for a given input: the network *learns* the correct weights for a task.

There are two types of learning: supervised and unsupervised [Pao 1989]. Supervised learning involves presenting training examples to the network and adjusting the weights so that the outputs match given, desired outputs. If the weights can be adjusted so that the network gives the desired output for all training examples, learning is successful. Unsupervised learning involves presenting a series of training examples and letting the network classify these into different classes. The weights are adjusted to classify the examples accordingly. A measure of similarity is needed, which is often the Euclidean distance between input examples [Rumelhart and McClelland 1986]. Examples that are close in terms of Euclidean distance are clustered as one class, and an output node is assigned to that class. The network forms as many classes as needed to classify all the training examples. In the main, supervised learning is used for engineering problems, as for most problems the desired outputs for the training examples are known.

Once a network has been trained to correctly classify the training examples, it may correctly classify other examples. A network has *generalised* when it successfully analyses examples other than the training examples, and it has *specialised* when it only performs successfully on the training examples.

Neural networks have been used to model human cognitive functions [Rumelhart and McClelland 1986] However, neural networks are more commonly used for pattern recognition and other ill-defined problems [Lippmann 1989]. They have been shown to be effective for many problems, and this has spurred is a lot of research into the use of neural networks as an analysis tool [Widrow and Lehr 1990].

4.6.1.2 Neural Network Design

There are many different types of neural networks, and each type has a number of variables. This section describes the characteristics of neural networks, focusing on the layered feedforward architecture.

Transfer function

The nodes are threshold units that sum inputs, and passing the sum through a *transfer function* to give an output (*f* equation (4.8)). Initially a binary function was used, but non-binary functions are the norm [Widrow et al. 1988]. Biologically, a smooth function is believed to be more plausible because neurons do not behave in a binary manner. The output level of the node is usually constrained between 0.0 and 1.0, with some groups using -1.0 and 1.0. The most common function is the sigmoid [Burr 1988, Felten et al. 1990, Scalero and Tepedelenliogu 1992]:

$$f(x) = \frac{1}{1 - \exp(-x)} \tag{4.9}$$

Another common function is the hyperbolic tangent [Ruck et al. 1990]:

$$f(x) = \tanh(x) \tag{4.10}$$

The hyperbolic tangent was found by one group to allow faster learning than a sigmoid transfer function, with no accuracy difference once trained [Sorsa et al. 1991]. However, the sigmoid is the most commonly used transfer function [Lippmann 1987].

Input and output

Neural networks usually analyse features as opposed to raw data. Raw data often contains redundant information that causes training problems, and requires a larger network. However, extracting features requires prior knowledge about the patterns to be analysed. The number of features or data samples to be used as external inputs to a neural network must be decided by the user according to the problem.

The standard external output is one node for each class or feature to be recognised. The number of outputs is typically quite small. A pattern is presented to the input nodes, and a high (close to maximum) output at an output node indicates that the input pattern belongs to that class, whereas a low (close to minimum) output means that the pattern is not a member of that class.

Architecture

The architecture is partly defined by the number of external inputs and outputs. The number of hidden layers, the number of nodes in each layer, and the connections between nodes are variable. A larger network is capable of solving more complex problems, and may be easier to train. The complexity of a network is related to the number of weighted connections [Baum and Haussler 1989]. The number of training examples needed increases with the number of weights and the number of nodes. Therefore, for more complex problems larger networks are needed, and these require more computational power and more storage.

A smaller network, if trained successfully, is more likely to generalise. However, it has been shown that an exponential increase in the number of nodes leads to a linear decrease in training time [Rumelhart and McClelland 1986]. Hence, the aim is to have a network large enough to train quickly, large enough to handle the complexity of the problem, but small enough to generalise [Pal and Mitra 1992].

Whilst the architecture is usually determined before training, some training strategies include determining the architecture size [Brent 1991, Bello 1992]. There are methods for determining the minimum number of hidden nodes required, but these require statistical knowledge about the problem [Fogel 1991].

Most applications use one or two hidden layers; a three-layer network is in theory capable of performing tasks of any complexity, but multiple layers can improve performance [Nilsson 1965, Lippmann 1987]. The outputs of each node in one layer are used as inputs to every node in the following layer. Too many layers can impede performance, with four-layer networks either failing completely to train, or taking longer to train and not generalising [Pao 1989, Sorsa, et al. 1991]. De Villiers and Barnard [1992] also found that training performance with three layers was better than with four layers. Once trained, they found no difference in performance between three and four layer networks [de Villiers and Barnard 1992]. However, Pal found that three-hidden-layer networks performed best in terms of recognition and training time [Pal and Mitra 1992]. Another point is that a two-hidden-layer network is easier to train if number of nodes is balanced [de Villiers and Barnard 1992]. As most studies show no advantage for more than four hidden layers, networks with one, two and three hidden layers are common.

Criterion function

A criterion function is a measure of the performance of the neural network. It is a penalty function [Beveridge and Schechter 1970], which has large values for mis-classified examples. For a network to train successfully, the criterion function needs to be reduced to below a threshold.

The simplest criterion function is the Bayes criterion function, where a division is made between classes as a function of the pattern and the weights, and the criterion function is either 0 or 1 (for a two class problem) [Barnard and Casasent 1989]. The criterion function used in the perceptron is the error between the actual and desired output [Widrow and Lehr 1990]. An error is associated with each pattern, and the total error is the average error over all patterns. The most common criterion function is the mean square error (MSE). The MSE may be calculated either for each pattern or averaged over all patterns [Widrow, et al. 1988, Qin et al. 1992, Scalero and Tepedelenliogu 1992]. Entropy has also been used as a criterion function, but does not offer an improvement over the MSE [Brent 1991, Pal and Mitra 1992, Takagi et al. 1992]. There are different criterion functions that have been used for specific problems [Barnard and Casasent 1989], but the MSE is used for most applications.

4.6.1.3 Training strategies

Training a neural network is an optimisation problem: select the weights in the network such that the criterion function is minimised. There are many ways to solve this problem, just as there are many methods of optimisation. This section describes the characteristics of training a feedforward neural network, and some of the different options available.

Training examples

A training example consists of inputs and associated desired outputs. The number of training examples required for successful training depends on the complexity of the problem, and on the quality of the training examples. It has been shown theoretically that for a given maximum error E_{net} and a network with W weights and N nodes, the number of training samples required is in the order of W/E_{net} to $W/E_{net}\log(N/E_{net})$ [Baum and Haussler 1989]. Another guide is that the number of training examples required is in the order of M.min(n,d), where M is the number of classes, n the number of input features and d the number of first hidden layer nodes [Mehrotra et al. 1991]. However, it is important that the examples be as different as possible: obviously, two almost identical examples carry little more information than a single example. It is important to have as many *boundary* examples as possible, where boundary examples are examples that belong to one class but that are similar to examples in another class [Mehrotra, et al. 1991].

Finally, the desired output levels for training examples are not set to exactly 1.0 or 0.0, as these extreme values cannot be reached using normal training methods [Pao 1989]. Instead, values such as 0.9 or 0.1 are used.

Weight initialisation

As learning is a search in weight space, the initial weights have a strong influence on training [Guo and Gelfand 1991]. If all the weights are initialised to the same value, then the network cannot learn, and therefore weights are typically initialised to random values within a range [Pao 1989]. If the initial weights are close to the optimum, training will be fast; conversely, if the initial weights are not close to optimum values, learning is likely to be slow, and local minima may stall the optimisation and cause the learning to be unsuccessful. Therefore, for a given problem, training is repeated for several different random weight initialisations. Examples of the ranges within which random weights are initialised include: -1.0 to 1.0 [Pal and Mitra 1992]; -0.3 to 0.3 [Takagi, et al. 1992]; -0.5 to 0.5 [Scalero and Tepedelenliogu 1992]; and -0.1 to 0.1 [Yu and Cheng 1990].

Batch or pattern learning

The search in weight space is aiming to minimise the criterion function, which is a function of the error between actual and desired outputs. The error may be for each pattern or averaged across all training patterns; the former is *pattern learning* and the later *batch learning*. With batch learning, the error over all patterns is calculated and the weights are adjusted. With pattern learning, the error after each example is calculated and the weights are adjusted before the next training pattern is presented. Batch learning is a true gradient search, whereas pattern learning is more biologically plausible, and it also potentially allows for faster training. Batch learning has been found to lead to better performance than pattern learning, although performance is similar in some instances [Vogl et al. 1988, Qin, et al. 1992]. Another method combines batch and pattern learning by using *local batches*, which involves calculating the error over groups of training examples. Local batch learning is appropriate if there are many training examples [Bello 1992].

Backpropagation

Backpropagation is an optimisation method for adjusting the weights in order to successfully train a neural network. Backpropagation is the original method of training used for standard classification problems, and is still the benchmark today [Rumelhart and McClelland 1986].

Training using backpropagation is conceptually simple: an error is associated with each node based on the difference between the actual and a desired output, and then the weights into that node are adjusted to reduce the error [Rumelhart and McClelland 1986]. Thus, an error needs to be associated with each node. This is simple for the output nodes, as the error is the difference between desired and actual response. The backpropagation algorithm then "back-propagates" errors from the output layer to previous layers, thereby determining an error for every other node in the network.

During training, examples are presented to the input layer, propagated forward to the output layer, and the actual responses are compared to the desired responses. The difference is the error, which is backpropagated through all nodes in the network so that each node has an associated error. The value of the weightings of the connections into each node are adjusted to reduce the error at that node. The training examples are presented again, and the process is repeated. The process is continued until the value of the criterion function is acceptably small, or training fails (no reduction in the criterion function value after a large number of training iterations) [Rumelhart and McClelland 1986].

Training with the above method is slow, requiring many iterations to train even simple tasks. To increase convergence, two parameters are introduced: *learning rate* and *momentum*. The reasoning for these parameters is that the solution converges on average, but there are many fluctuations. If the weights are decreasing on average, the solution would be reached faster if the fluctuations in weight changes at each iteration were smoothed out. Both parameters are used to smooth the weight changes. Learning rate is the proportion by which the weights are changed at each iteration. The momentum is the proportion of the previous weight change to include in the current update, and is based on the idea that the weights should be changing in approximately the same direction.

Values for learning rate and momentum are determined according to each problem. There is no one optimum value, and it has been shown that, for any problem, the optimum value changes at each iteration [Kuan and Hornik 1991]. If the learning rate is kept constant, convergence may be poor [Kuan and Hornik 1991], but if learning rate is adapted, further computations are required. One method is based on initially changing the weights in large steps, and then reducing the size of the step changes [Pal and Mitra 1992]. This method is based on the idea that during the initial iterations, the weights move quickly towards the general solution region, and as the solution narrows down to an optimum, smaller and smaller step sizes are required. The approach is to start with a high value of learning rate and a low value of momentum, and gradually decrease and increase these respectively during training [Pal and Mitra 1992]. The learning rate may be varied according to the average gradient, usually starting high then reducing [Park et al. 1992] However, even though learning rate and momentum can be adjusted at each iteration, the methods to do so are *ad hoc*, and optimality and robustness are not assured [Vogl, et al. 1988].

Learning rate is usually in the range of 0.7 to 0.9; Kollias and Anastassiou showed that for a range of learning rates from 0.1 to 0.9, 0.9 gave the best performance [Kollias and Anastassiou 1989]. Qin et al. found that a learning rate of above 0.14 cause oscillation problems, but this was only tested on one simple problem [Qin, et al. 1992]. Reyneri and Filippi used a very small learning rate of 0.01 to 0.05, which was required because pattern learning was used rather than

batch learning [Reyneri and Filippi 1990]. Momentum is usually in the range of 0.5 to 0.7 [Rumelhart and McClelland 1986], but few studies have found an optimum value.

Other training techniques

Numerical optimisation techniques are well researched, and have been applied with success to the training of neural networks. One common method is conjugate-gradient descent [Fletcher and Reeves 1964, Powell 1977, de Villiers and Barnard 1992]. This was found to perform better than steepest-descent and almost as well as BFGS variable-metric, a more complex technique requiring more computation and storage [Barnard 1992]. Brent also achieved good performance using this method [Brent 1991]. These results were confirmed by Yu and Cheng [1990], who also found conjugate-gradient better than steepest descent whilst not quite as good as a Gauss-Newton method. Another quasi-Newton method was found to have much improved performance compared to backpropagation [Bello 1992]. As well as training faster, these methods also have fewer parameters, such as learning rate and momentum, that need to be tuned.

Some training algorithms involve building the network architecture as well as determining the weights. One such is based on decision trees to construct architecture as well as determine weights; a series of optimisations is performed to define boundaries of hyperplanes, so for many inputs, these are in themselves large optimisation problems [Brent 1991]. Another approach is to optimise the criterion function with respect to the inputs to nodes rather than weights. This is significantly more complicated than backpropagation, but has the advantage of a two-fold increase in training speed and less sensitivity to initial weights. Large matrix manipulations are needed for large problems with many training examples [Qin, et al. 1992, Scalero and Tepedelenliogu 1992]. Recursive least squares optimisation has been used, for a two times improvement in training time over backpropagation; this uses past learning rates and adjusts them each iteration [Azimi-Sadjadi and Liou 1992]. Similarly an adaptive least squares method is better than backpropagation but still needs learning rate tuning [Kollias and Anastassiou 1989]. For large scale problems, small neural networks have been used to extract features, and the outputs of these small networks were used as inputs to a second-stage network [Giles and Maxwell 1987, Felten, et al. 1990]; this method requires some *a priori* knowledge and a well-defined problem. Less successful methods have been tried, such as training based on a Kalman filter [Iiguni and Sakai 1992]: this did not reduce computational time, even though iterations were reduced.

In general, these methods are more complex than backpropagation, and the added complexity is only justified where backpropagation is not effective. Hence, if a network can be satisfactorily trained using backpropagation, there is little need for other optimisation methods.

4.6.2 Implementation of Neural Network Analyses

A variety of networks were implemented and tested on simple problems. The standard backpropagation training and feedforward neural network architecture was used [Rumelhart and McClelland 1986]. All networks had at least one hidden layer, as this has been shown to be necessary in order to solve non-linear problems [Nilsson 1965, Baum and Haussler 1989]. It has been shown that for an exponential increase in the number of nodes, there is a linear decrease in the training time, up to the point where the network no longer generalises [Rumelhart and McClelland 1986]. In other words a smaller network, if trained successfully, is more likely to

generalise [Baum and Haussler 1989]. Thus, the number of nodes is a balance between being large enough to handle the complexity of the problem and train within a reasonable time, and being small enough to generalise [Pal and Mitra 1992]. When two or more hidden layers were used, the number of nodes in each layer were kept within 20% of each other, as such architectures have been shown to train more successfully [de Villiers and Barnard 1992].

There are training algorithms other than backpropagation, although many rely on extra information regarding the problem that the neural network is being trained to solve [Iiguni et al. 1992]. The main advantage of other training algorithms over the backpropagation algorithm is the speed with which the neural networks are trained [Brent 1991, Bello 1992, de Villiers and Barnard 1992]. However, many methods make assumptions about the nature of the data or problem, whereas backpropagation is a general method that is based on very few assumptions regarding the problem. Given that a neural network has trained successfully, the performance is independent of the training method [Rumelhart and McClelland 1986]. Backpropagation is the reference standard for all new training algorithms, and is therefore used in this research.

A variety of neural networks were tested. However, before being applied to the more complex problem of apnoea detection, they were tested on more simple problems. Three simple tasks were designed:

- exclusive-OR, two inputs and one output;
- parity 3, three inputs and one output;
- recognising a pattern within a 4x4 image.

For these three problems, various parameters were tested and the values that lead to reliable and fast training were recorded. If the training was successful, the training time was shorter if the initial weights were randomly initialised with a wide range of values, as opposed to initialising weights using a small range of values. Therefore, a network initialised with weights between -5.0 and 5.0 trained faster than a network with initial weights between -1.0 and 1.0. However, initialising weights between -5.0 and 5.0 resulted in networks failing to learn in 30-40% of trials. After testing a variety of initialisation ranges, the -1.5 to 1.5 range was the largest initialisation range that consistently trained successfully, and was selected as an initialisation range that allowed for both consistent and fast training. Networks with more than three hidden layers were not considered. There are few examples of these deeper architectures being used, and evidence to suggest that networks with less than four layers train faster and more successfully than networks with four or more hidden layers [de Villiers and Barnard 1992, Pal and Mitra 1992]. Summarising, parameters that allowed for fast and consistent training were:

- sigmoid transfer function;
- the backpropagation training algorithm;
- random weight initialisation from -1.5 to 1.5;
- momentum in the range 0.5-0.7;
- learning rate in the range 0.7-0.9;
- one or two hidden layers;
- batch learning;
- desired outputs set at 0.1 for non-detection and 0.9 for detection.

These parameters are described in detail in the neural network literature [Rumelhart and McClelland 1986]. The above parameter settings allowed for fast and consistent training in all three of the simple problems listed above.

An initial approach to apnoea detection was implemented using a neural network to directly analyse a Graseby breathing signal. Sturman [1991] had used a five second, 50 node input with two hidden layers of 50 nodes each. The aim was to replicate the same network but apply it to a wider number of recordings and training events. The network and training parameters were based on the trials with the three simple problems, and were set as above.

The neural network was trained to detect five second apnoeas, which was the reason for using five second segments of breathing as input to the network, resulting in 50 input nodes (the breathing was sampled at 10Hz—see Section 2.2.1). One output node indicated whether an apnoea had occurred or not. Equal numbers of apnoea and non-apnoea events were used in the training sets. Sturman had previously used a 50-50-50-1 layer network, which generalised to a small test set [Sturman 1991]. However, in replicating this network and testing on larger data sets, it was found to perform poorly. Whatever training and network parameter values were used, no network trained successfully. The reason is possibly due to the fact that the sample points, which are the inputs, have low discriminating power. The network had only subtle differences between the inputs to use for discrimination.

In conclusion, neural networks do not appear to be suited to analysing raw breathing data. With Sturman's work, it is possible that the network did not generalise to more than the test data, as the test and training data came from less than an hour of one recording [Sturman 1991]. In line with other uses, a neural network approach may succeed when used in conjunction with some feature extraction algorithm or other data reduction method.

4.7 Discussion and Conclusions

The objectives of this research are to design specific mathematical descriptions of signals that correspond to apnoeas, and using these to develop a reliable detection system using the Graseby signal. This section revises the findings of this chapter in relation to these objectives.

The measures discussed in Section 4.2 are mechanisms by which the performance of any algorithm can be evaluated. For an individual algorithm, the numbers of false positives and false negatives give an indication of the performance (Section 4.2.2). Different algorithms need to be compared so that the best can be selected, and a performance measure is used in the form of a penalty function (equation (4.1)). The penalty function is a mathematical description of human expert opinion. There is no absolute performance measure, but nevertheless, the penalty function does provide a means of quantifying performance. Using this penalty function to evaluate apnoea detection algorithms is more rigorous than asking for experts' opinions each time a new algorithm has been tested. The results can be defined using a confidence interval that gives an indication of the accuracy of the performance measure, and a confidence interval is especially important given the uncertainty levels associated with the reference sets (Section 4.2.4). Thus, there are performance measures that can be used as a basis for developing detection algorithms.

Existing methods of breathing signal analysis successfully perform tasks such as segmenting breathing into breaths or measuring median breathing rate, as described in Section 4.2. However, an accurate detection algorithm requires higher levels of accuracy than available with existing methods of analysing breathing. In order to improve apnoea detection performance relative to the general breathing analysis algorithms, apnoea detection algorithms need to take into account events that are represented by flat regions in the breathing signal but that are not apnoeas. Mason states that the variability of the breathing signal waveforms *...pose[s] enormous problems for any detection routines* [Mason, et al. 1974]. In the field of evaluating EEG's, false positives that are similar to epileptic spikes or other features being searched for have been classified as "interesting false positives" [Gotman 1990]. In the context of breathing, "interesting false positives" could be events that are not apnoeas, but that are abnormal breathing behaviour. Taking these abnormal events into account is what would distinguish apnoea detection algorithms from general breathing analysis algorithms.

The traditional peak-to-peak detection algorithm for analysing breathing performs poorly when applied to the Graseby signal for apnoea detection, even though it successfully calculates the breathing rate on the same signal. However, the peak-to-peak algorithm does detect the majority of breathing pauses, and could possibly be used as a method of preprocessing the breathing signal before further analysis. The neural network approach failed, mainly because the raw data is not suitable for direct input to a neural network. It is still possible that a neural network could be used as part of a detection system.

Approaches to apnoea detection that do not appear promising involve mapping the breathing signal to another domain; the examples discussed in Section 4.5 were frequency domain mappings. Representing the signal in another domain means that it then needs to be interpreted in that domain, and in particular, the features that correspond to apnoeas need to be defined in the new domain. Within the time domain, an apnoea is essentially well defined as a flat region, and it is the details of the flat region that experts use to distinguish between a true cessation of breathing and other events. A problem with methods that first transform data to another domain is that existing transformations have not been designed specifically to enhance features of apnoeas, and may filter important details. Thus, what is necessary with any mapping is retaining and enhancing important details, which in the case of apnoea detection are flat regions and geometric details of flat regions (see Section 3.3). It is these details that the experts use to discriminate between apnoea and non-apnoea events, and hence these details that any detection algorithm must measure or enhance.

Many signals have redundant data in terms recognising a pattern and so if the data can be reduced, the decision-making task may be simplified. In terms of apnoea detection, given a 30 second segment of data with 300 data points, there are some redundant data. If measures of features that distinguish between apnoea and non-apnoea events can be calculated, then classification could be performed using a small number of measures, as opposed to 300 data points. There are a variety of methods that rely on calculating features or parameters of a signal, and performing classification based on measures of the features [Siegwart, et al. 1995, Wilks and English 1995]. As explained in Section 4.6, a neural networks are typically used with a small

number of salient inputs, as opposed to analysing the raw data directly, and this approach could also be applied to apnoea detection.

Concluding, two considerations can be drawn from the study of various approaches to analysing breathing. Firstly, traditional methods are not sufficient to enable accurate definition and detection algorithms to be developed. The methods that have been used to analyse breathing point to possible new approaches, but in essence are not designed for apnoea detection. This problem was summarised several years ago: *...computerized analyses of overnight respiration recordings can be unreliable or misleading* [Miles, et al. 1989]. Thus, to meet the objectives of this research, new a approach is required.

Secondly, a promising general approach is to reduce the raw breathing signal data to a number of measures of features that distinguish between apnoea and non-apnoea. In other words, measures could be developed that measure the information specific to the geometric features of a breathing signal that represent apnoeas. Any classification as apnoea or non-apnoea would then be based on a small number of measures compared to the raw data. By splitting the detection process into feature extraction and classification, the problem may be simplified.

Chapter 5

A Statistical Method of Apnoea Detection: Development and Application

This chapter describes a statistical method of apnoea detection that is in current clinical use, as part of the BabyLog system. The method has been refined and had parameters optimised to maximise performance. Presented in this chapter is an example of the use of the method as a part of the analysis of a large volume of data collected during a study of normal infants. Because the clinical findings of the study relate to more than just apnoeas, all the analysis techniques used in the study are presented. Thus, the study provides an example of the context of apnoea detection.

5.1 A Statistical Method of Apnoea Detection

This section presents a statistical method for detecting apnoeas from a breathing signal. This method searches for flat regions of signal, and therefore relates to the manner in which experts define an apnoea signal (see Section 4.3), and some previous apnoea detection methods [Bruckert, et al. 1982, Laxminarayan, et al. 1983, Rakowski, et al. 1986]. A predecessor upon which the current algorithm is based was in clinical use for several years, but was not tested and no evaluation of performance had been performed [Dove, et al. 1990]. The original algorithm simply detected flat regions and classified these as apnoeas, and the algorithm was not considered reliable by clinicians. This original algorithm was refined and optimised, resulting in the current algorithm that is in clinical use, and that is presented in this section.

The algorithm is described as a statistical algorithm because it measures statistical properties of a breathing signal to detect apnoeas. The algorithm also measures other properties, but is referred to as a statistical algorithm to distinguish it from other algorithms presented in Chapter 6. Peirano et al. [1988] also used statistical measures to analyse a breathing signal. Specifically, Peirano's algorithm measures the standard deviation of a breathing signal, and detects flat regions using this measure. Bruckert et al. [1982] use the derivative of the signal, but this could be more sensitive to minor fluctuations as seen in Figures 3.1 and 3.2. As shown in Chapter 3, apnoeas are represented within a breathing signal as flat regions. Standard deviation is a measure of flatness: if the standard deviation of a signal drops then the signal has "flattened off." Having detected flat regions corresponding to possible pauses in breathing, the durations of these pauses are measured by detecting the start and end times of the pauses.

5.1.1 Detection of Flat Regions

The first stage of the apnoea detection method is detecting flat regions which, as explained in Chapter 3, correspond to possible pauses in breathing. An algorithm, which was developed for the original BabyLog system, was used to detect flat regions and hence possible pauses in breathing [Dove, et al. 1990]. Figure 5.1 illustrates the steps of this algorithm.

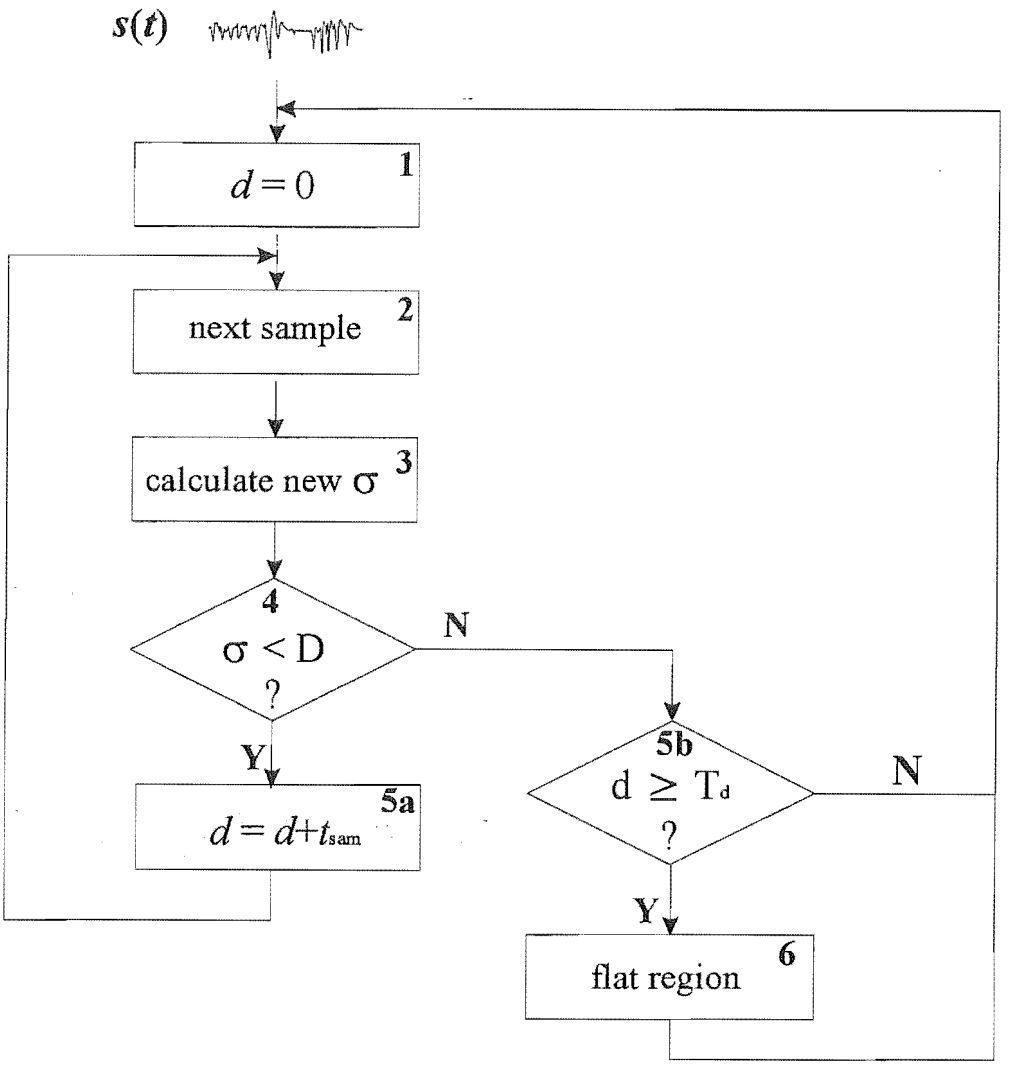

Figure 5.1 Statistical method of detecting flat regions: the standard deviation σ of a sampled breathing signal $s(t)$ is measured about subsequent samples. If σ is less than a threshold D for a time d, and d is at least T_d seconds, then a flat region F is recorded.

Before commencing, the time t_i of the current window of data being analysed is set to the beginning of the recording ($t_i = 0$). Referring to Figure 5.1, in step 1 the flat region duration d is initialised, and in step 2, the time t_i is incremented to the time of the subsequent sample:

$$t_i = t_i + t_{sam} \tag{5.1}$$

where t_{sam} seconds is the time between samples. (The sampling rate for the Graseby signal is 10Hz so $t_{sam} = 0.1$ seconds—see Section 2.2.) In step 3, the standard deviation σ is defined as the standard deviation of the breathing signal $s(t)$ over a window L where $t - \frac{L}{2} \le t \le t + \frac{L}{2}$. As is standard practice when calculating the standard deviation of a discrete signal [Neter, et al. 1978], the window length L is an odd multiple of t_{sam}:

$$L = n\,t_{sam} \tag{5.2}$$

where the integer n is odd and positive.

The standard deviation σ is compared to a standard deviation threshold D in step 4. This threshold is similar to the derivative threshold used by Bruckert et al. [1982]. If σ is less than D, the duration of the flat region d is incremented by t_{sam} (step 2) and the next window of data

analysed (step 3). Steps 2 to 5a are repeated until a time t_i is found for which $\sigma \geq D$; d is then compared to the duration threshold T_d, the duration in seconds for which the signal must be flat before it is recorded as a flat region (step 5b). If $d < T_d$, then the event is rejected as a flat region and the next region of $s(t)$ analysed. If $d \geq T_d$ then a flat region F is recorded (step 6); F is defined as the signal between a start time t_s and an end time $t_s + d$, where d is the duration:

$$F = [t_s, t_s + d] \qquad (5.3)$$
$$t_s = t_i - d$$

Once the flat region F is recorded, d is set to zero (step 1) and the process is repeated until the end of the recording.

There are three parameters that influence the performance of the algorithm—window length L, standard deviation threshold D, and flat region minimum duration T_d—and these parameters can be adjusted according to the nature of the events being detected and the nature of the recordings. The effects of each of the parameters on the detection performance is explained, and the values used are presented in Section 5.1.2.3.

Window Length L

The window length L is the duration in seconds of the window of breathing data over which the standard deviation is calculated. Figure 5.2 illustrates a 30 second segment of a breathing signal, and three standard deviations of the signal calculated using various window lengths; the segment of breathing contains two flat regions, the longer of which corresponds to an apnoea.

For all three window lengths, the standard deviation is low during the apnoea region (seen in Figure 5.2 between approximately 15 and 19 seconds of the record) and remains at a constant low level. However, the longer window lengths result in shortened regions of low standard

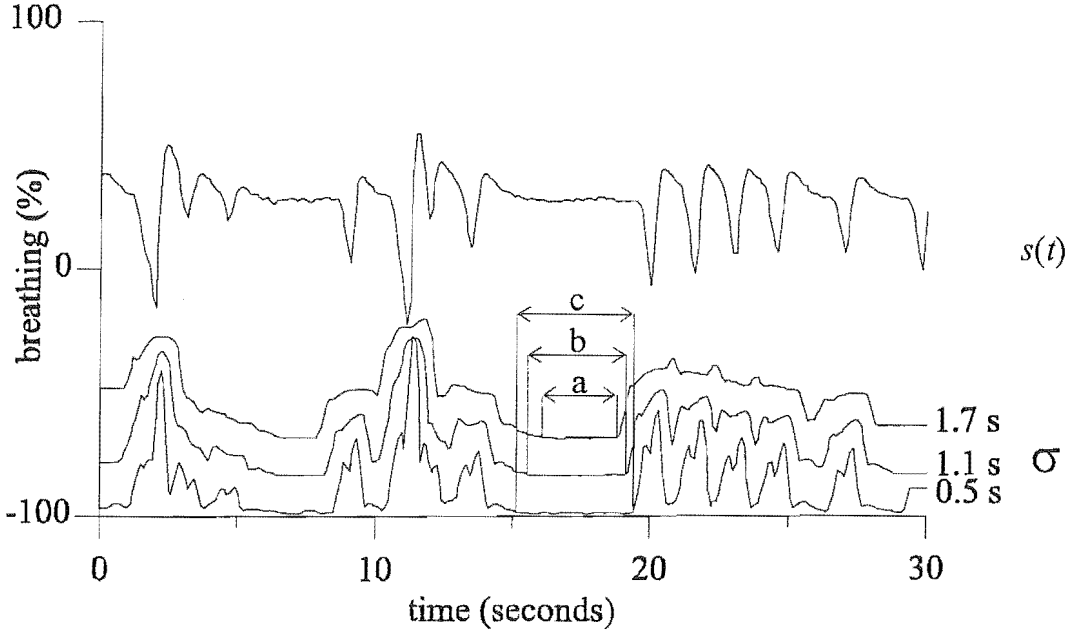

Figure 5.2 Standard deviations σ of a 30 second segment of breathing signal $s(t)$ which contains two flat regions, the longer of which is an apnoea. The standard deviations are calculated for varying window lengths; the y-axis does not apply to the standard deviations as they are scaled and shifted for display purposes. Note that the standard deviation calculated over a shorter window is low for a longer period of time c, whereas the standard deviation calculated over a longer window is low for a shorter period of time a.

deviation. Considering the standard deviations that correspond to the flat region of the apnoea in Figure 5.2, for L = 0.5 seconds, the standard deviation remains low for approximately 0.5 seconds longer at each end of the region of low standard deviation than the standard deviation calculated with L = 1.7 seconds. Thus, standard deviations calculated with shorter window lengths emphasise the flat regions of the breathing signal.

For all three window lengths, individual breaths in the breathing signal are seen as oscillations in the standard deviation. However, the size of the oscillations decreases as the window length increases. For a 1.7 second window, the oscillations in the standard deviation corresponding breaths are small, appearing as bumps—for example, in Figure 5.2 from time 20 to 25 seconds. The standard deviation calculated with shorter window lengths does not smooth out breaths as much as the standard deviation calculated with longer window lengths, and hence the standard deviation for shorter window lengths oscillates more. The standard deviation calculated using a window length of 0.5 seconds, shown in Figure 5.2, drops to low levels almost each breath during the period of breathing from approximately 20 to 25 seconds in Figure 5.2. Therefore, if longer window lengths are used, the standard deviation during breathing remains consistently higher, and can therefore be more distinct from the standard deviation during a flat region.

Thus L must be short enough that flat regions are detected, and long enough that the regions of breathing are distinct from the flat regions.

Standard Deviation Threshold D

The standard deviation threshold D is the value of standard deviation below which a region is considered flat, representing a pause in breathing. The smaller the value of D, the more certain that detected flat regions represent cessations in breathing. The larger the value of D, the greater the probability of detecting flat regions that do not correspond to pauses in breathing, but the greater the probability of detecting all apnoeas. The signal amplitude in each recording varies, and oscillations due to cardiac movement often appear on the flat regions, as seen in Section 3.3. Hence, a relatively high D is required to detect all flat regions corresponding to apnoeas, resulting in a high number of false detections.

Duration Threshold T_d

The duration threshold T_d is the minimum duration of a flat region. The smaller the threshold, the more flat regions are detected so fewer apnoeas are missed, but more false events are detected. Using a related algorithm, a threshold of 0.5 seconds has been found to be the maximum length for detecting three second apnoeas [Bruckert, et al. 1982]. Conversely, as T_d increases, there are fewer flat regions detected and therefore fewer false detections, but there are more missed apnoeas. This is especially so with high amplitude signals, where the decay from a high amplitude peak or trough may last several seconds, reducing the duration of the flat region of the signal. In fact, the flat region of a five second apnoea may be as short as two seconds (see Section 3.3). In order for the algorithm to be sensitive to the minimum duration events, T_d must be set to less than the minimum duration.

5.1.2 Measurement of Duration

This section describes algorithms for detecting the start and end times of pauses in breathing, and hence calculating the durations of pauses. Flat regions represent possible pauses in breathing, and the

previous section describes how flat regions are detected and recorded as a start time t_s and a duration d. However, t_s is not the start time of the breathing pause and d is not the duration of the entire breathing pause, and hence a method is required to detect the start and end points of a breathing pause according to the apnoea definition in Section 3.2.1.

A pause in breathing is represented by a flat region in a discrete breathing signal $s(t)$, and a flat region is recorded as a start time t_s and a duration d; hence the end time t_e of the flat region is defined:

$$t_e = t_s + d \tag{5.4}$$

The start of the pause in breathing is taken to be the last significant peak or trough before the beginning of the flat region, as defined in Section 3.2.1. The end of the pause in breathing is the end of the flat region, corresponding to the next inhalation or exhalation. Algorithms were developed to detect the start and end points of actual pauses in breathing, defined respectively as t_{start} and t_{end}.

5.1.2.1 Start Time Detection

Based on the definition in Section 3.2.1, the start time t_{start} is the time of the first significant peak or trough prior to the flat region, and t_{start} is calculated using the algorithm illustrated in Figure 5.3. A typical breathing signal has many peaks and troughs, but only those which correspond to the end of breaths (the end of inhalations or exhalations) are of interest. Steps 1 to 4 detect peaks and troughs in the signal, and steps 5 to 9 check whether each peak or trough is valid.

Starting from t_s, samples at times t_j are tested in order to determine whether they fit the criteria that define the start time. The times considered are prior to t_s, as the start of an apnoea is by definition prior to the start of the flat region. The limit T_{pw} is the maximum time prior to t_s that the start point can be detected. If no valid peak or trough has been found within T_{pw} of t_s (step 2), then the default value for the start of the pause is the start of the flat region, and is set in step 3a as shown in Figure 5.3.

Peaks and troughs are detected in step 4; given a possible start time t_j, a peak is detected if the value of $s(t_j)$ is greater than surrounding values:

$$s\left(t_j - t_{next}\right) > s\left(t_j\right) < s\left(t_j + t_{sam}\right) \tag{5.5}$$

or a trough is detected if the value of $s(t_j)$ is less than surrounding values:

$$s\left(t_j - t_{next}\right) > s\left(t_j\right) < s\left(t_j + t_{sam}\right) \tag{5.6}$$

where t_{next} is the time of the next sample prior to t_s that is not of equal value, and is defined by the following:

$$t_{next} = \mathrm{n}t_{sam} \tag{5.7}$$

where

$$\mathrm{n} = \min\left\{\mathrm{i}: s\left(t_j\right) \neq s\left(t_j - \mathrm{i}t_{sam}\right)\right\} \tag{5.8}$$

As most adjacent samples are not the same value, n is usually 1 and $t_{next} = t_{sam}$. When t_j is incremented in step 10, it is incremented to the time of the next sample not of equal value:

$$t_j = t_j + t_{next} \tag{5.9}$$

Thus, flat peaks and troughs are treated as occurring at one time, not over a period of time.

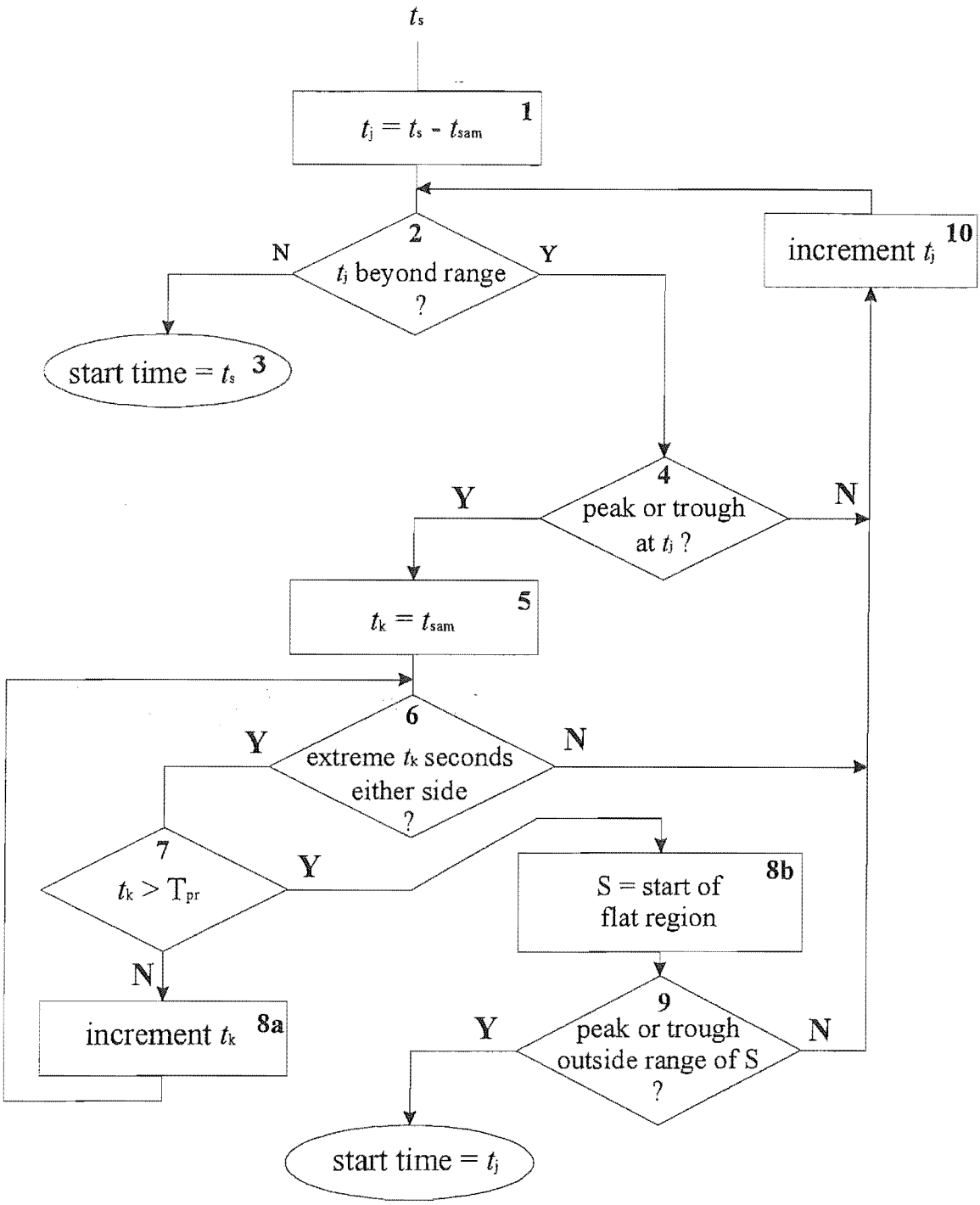

Figure 5.3 Detecting the start time of a pause in breathing, which is the time of the peak or trough prior to the flat region. The time t_{sam} is the time between samples, and the time t_s is the start time of the flat region. T_{pr} is a time threshold, and S a the set of samples that constitute the start of the flat region.

A check is performed in step 6 to ensure that the peak or trough is the extreme value for T_{pr} seconds before and after the peak or trough itself. Samples are checked one at time, with step 6 being repeated for different values of t_k until T_{pr} is reached (step 7). Steps 5, 6, 7 and 8a check for the following conditions, firstly for peaks:

$$s(t_j) > s(t_j + t_k) \qquad \forall\ t_k : -T_{pr} \le t_k \le T_{pr} \tag{5.10}$$

and secondly for troughs:

$$s(t_j) < s(t_j + t_k) \qquad \forall\, t_k : \; -T_{pr} \le t_k \le T_{pr} \tag{5.11}$$

If a detected peak does not meet the conditions specified in equation (5.10) or if a detected trough does not meet the conditions specified in equation (5.11), then the time t_j is rejected as a possible start time, and the next sample considered (step 10).

Finally, steps 8 and 9 test the amplitude of the peak or trough. If the amplitude is within the amplitude range of the flat region, the peak or trough is rejected; in other words, the amplitude of a valid peak or trough must respectively be greater than or less than the values of the signal in the flat region, which occur about a rest value (see Section 2.3.3). A peak or trough is considered in relation to the start of the flat region, and therefore, in step 8b, a set S of samples that constitute the start of the flat region is defined:

$$S = \{s(t) : t_s < t < t_s + L_p\} \tag{5.12}$$

where L_p is the duration of the portion of the flat region. In step 9, the peak or trough is considered valid if the following holds:

$$s(t_j) > \max\{S\} \quad \text{OR} \quad s(t_j) < \min\{S\} \tag{5.13}$$

If neither of these conditions hold, then the next sample is evaluated after incrementing t_j (step 10). Otherwise, the time of the peak or trough is defined to be the start time of the pause in breathing ($t_{start} = t_s$).

5.1.2.2 End Time Detection

An algorithm for locating the end point t_{end} of a pause in breathing is shown in Figure 5.4. Given a possible apnoea represented by a flat region with end time t_e, the end time of the apnoea or pause in breathing is defined as the time of the first significant deviation away from the flat region. The flat region, as detected using the method in Section 5.1.1, is defined by the region of low standard deviation of $s(t)$ which, as can be seen in Figure 5.2, is shorter than the pause in breathing represented by the breathing signal. Therefore, the end time of the breathing pause cannot be before the end time of the detected flat region ($t_{end} \ge t_e$). The end time t_e is the time after which the standard deviation of $s(t)$ about a window length L rises above the standard deviation threshold D; the end time of the window is defined as t_L:

$$t_L = t_e + \tfrac{L}{2} \tag{5.14}$$

Therefore, $s(t)$ after t_L must deviate from the flat region enough to raise the standard deviation above the threshold, and t_e is defined as being less than or equal to t_L; Thus:

$$t_e \le t_{end} \le t_L \tag{5.15}$$

Time t_j is initialised (step 1 in Figure 5.4) and the thresholds A_{max} and A_{min} are calculated (step 2), which are the maximum and minimum values respectively of the breathing signal $s(t)$ during the detected flat region:

$$\begin{aligned} A_{max} &= \max\{s(t) : t_s \le t \le t_e\} \\ A_{min} &= \min\{s(t) : t_s \le t \le t_e\} \end{aligned} \tag{5.16}$$

Each successive sample after t_e is checked by incrementing t_j (step 4b):

$$t_j = t_j + t_{sam} \tag{5.17}$$

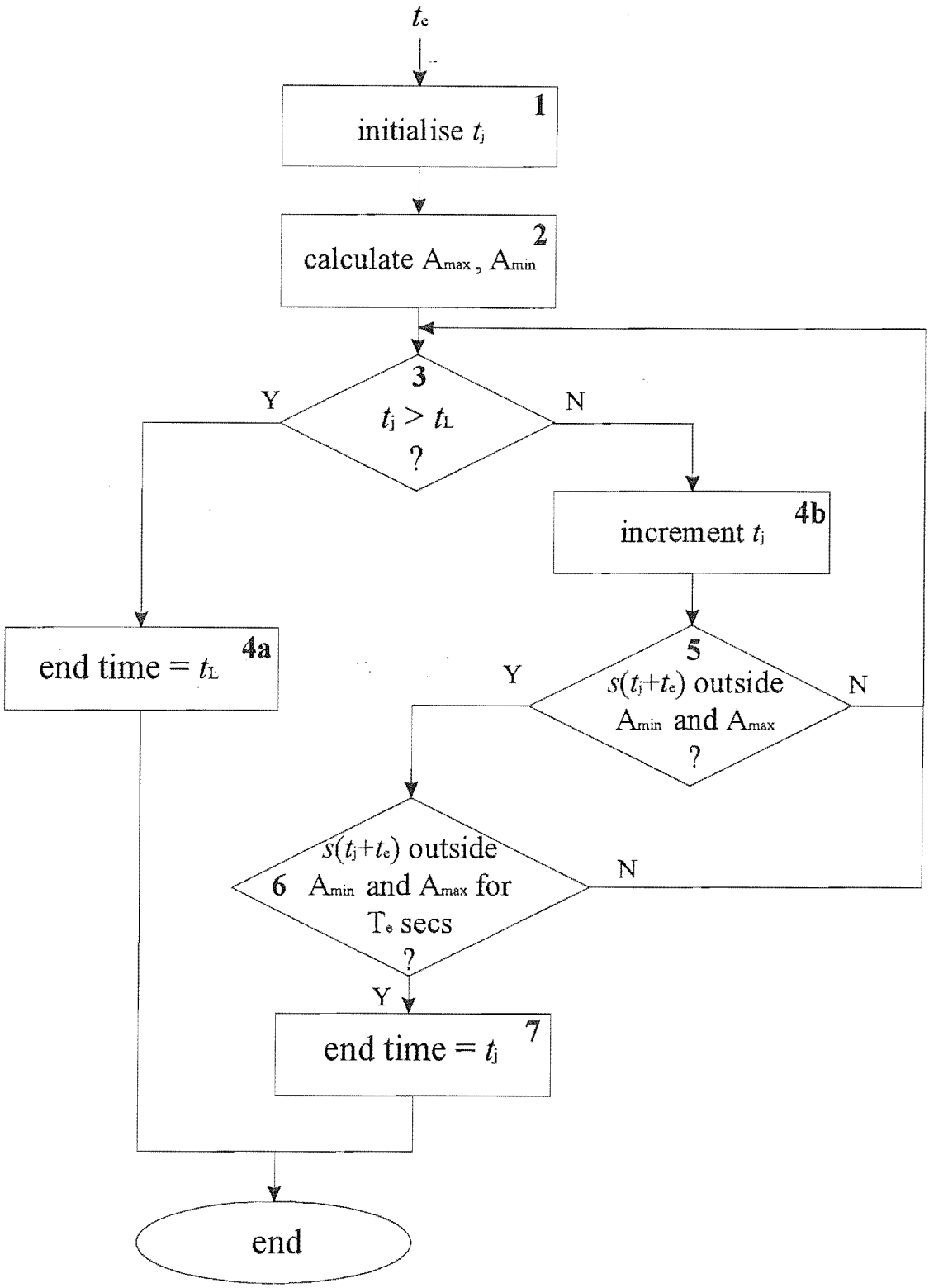

Figure 5.4 Detecting the end time of a pause in breathing, given a breathing signal $s(t)$ and a flat region ending at time t_e. The end time is between t_e and the time t_L of the end of the window in which σ was low (see Section 5.1.1).

where t_{sam} is the time between the samples of $s(t)$. Next, $s(t_j)$ is tested to ensure that it has deviated significantly—above A_{max} or below A_{min} (step 5), for T_e samples (step 6). The conditions are as follows:

$$\min\left\{s\left(t_e + t_j + n t_{sam}\right)\right\} > A_{max} \quad \text{OR} \quad \max\left\{s\left(t_e + t_j + n t_{sam}\right)\right\} < A_{min} \qquad (5.18)$$
$$\forall n = 1, 2, \ldots, T_e$$

where n and T_e are integers. If either of the conditions in (5.18) hold, then the time at which that sample occurs is classified as the end time t_{end} (step 7):

$$t_{end} = t_e + t_j \tag{5.19}$$

If there are no times at which either of the criteria in (5.18) are met, then the end time is set to the end of the window which was used to calculate the last time of low standard deviation (step 4a):

$$t_{end} = t_L \tag{5.20}$$

From the start and end times, the duration can be calculated. Pauses in breathing that are less than a minimum duration threshold are rejected, while the other events are recorded as apnoeas.

5.1.2.3 Optimum Duration Parameter Values

The algorithm has various parameters that need to be set, namely T_{pw}, T_{pr} and L_p for calculating the start time and T_e for calculating the end time. A subset of 50 apnoeas from B_1 was used to test and set these parameters. As the start times and durations of the apnoeas in B_1 were originally detected to the nearest second, three experts located the start and end times of these 50 apnoeas to the nearest sample (0.1 seconds). For these 50 apnoeas, the parameters T_{pw}, T_{pr} and L_p were set by adjusting them to minimise the Mean Square Error (MSE) between the detected start times t_{start} and the start times as determined by experts. Similarly, T_e was adjusted to minimise the Mean Square Error (MSE) between the detected end times t_{end} and the end times as determined by experts. The resulting values are shown in the table below:

Parameter	Value (seconds)
T_{pw}	2.4
T_{pr}	0.7
L_p	2.2
T_e	0.3

Table 5.1 Parameter values for calculating the start and end times of a pause in breathing.

Using these values, it was found that all detected start and end times were within two samples (0.2 seconds) of the start and end times of the 50 apnoeas as detected by human experts.

5.1.3 Results

An apnoea detection system was implemented, consisting of the algorithm for detecting flat regions (described in Section 5.1.1), the algorithms for detecting the start and end times (described in Section 5.1.2), and a final stage to calculate the duration and reject events that were less than the minimum duration. The performance of the system was measured relative to B_1 (a reference set of breathing signals from ten recordings and apnoeas based on three experts' opinions—see Section 3.2), and results were compared using different parameter settings.

The parameters T_{pw}, T_{pr} and L_p for detecting start and end times were set as shown in Table 5.1, but the parameters L, D and T_d were tuned to optimise a performance measure, P, as described in Section 4.2.3. Using different parameter settings, the breathing signals from B_1 were analysed, and then the performance measured as described in Sections 4.2.1 and 4.2.2. From the false positive and false negative rates, the penalty value, P, was calculated using (4.1). The lower the value of P, the better the performance.

The parameters were tuned across their entire feasible range. Some parameters have obvious limits: the smallest increment for T_d is one second, and the minimum value is zero seconds. As apnoeas of five seconds in duration are being detected, T_d must be five seconds or less, otherwise it will exclude the shorter apnoeas. Similarly, the window length L has a smallest increment of 0.1 seconds (one sample), must be greater than 0.3 seconds in order to be able to calculate the standard deviation, and must be less than 5.0 seconds in order to distinguish the region of low standard deviation during a five second apnoea. The standard deviation threshold has a lower limit of 0.0, with no minimum increment. Given that the range of the breathing signal is 200, the minimum increment was taken to be 1.0. The P values of the performances of the algorithm with the five best parameter settings are shown in Table 5.2, illustrating that L = 1.1 seconds, D = 2.0 and T_d = 2 seconds are the overall optimum settings.

Patient	Penalty Values for Different Parameter Settings L-D-T_d				
	1.1-2-2	**1.1-3-3**	**1.1-2-3**	**1.1-4-3**	**1.1-3-2**
1	49	47	32	65	70
2	63	65	151	37	47
3	57	48	65	58	65
4	49	54	40	109	85
5	102	136	64	157	163
6	114	160	84	185	179
7	93	69	159	64	79
8	80	69	40	95	105
9	63	47	71	72	75
10	107	101	90	123	135
Total	**777**	**796**	**796**	**965**	**1003**

Table 5.2 Penalty values calculated from (4.1) for the five best settings; the lower the value, the better the performance.

The overall optimum settings are not optimal for any of the individual recordings, reflecting the varied nature of the recordings. Overall, optimum performance was achieved with a standard deviation threshold D of 2.0, but for individual recordings the optimum D varied from 1.0 to over 5.0. For higher amplitude recordings, the system performed best using a higher value of D, 4.0 and above. A higher threshold was required because cardiac oscillations were amplified in high amplitude recordings, increasing the standard deviation during flat regions. On the other hand, for most recordings, a major cause of false detections was that regions of low amplitude breathing were detected as flat regions. A lower D can minimise the number of false detections due to low amplitude breathing. The recording for patient 6 contained several hours of low amplitude breathing, and required a D of 1.0 for optimum performance. On the other hand, patient 2 was reported by the experts to have no low amplitude breathing, and required D to be set at 5.0 for optimum performance. Thus, the optimum D for all recordings was a compromise which is reflected in the value for the overall optimum D of 2.0.

For the duration threshold T_d, two seconds is the overall optimum value, and two or three seconds are the optimum values for all individual recordings, regardless of L and D. If a T_d of

one second is used, then a large number of events (often several thousand per recording) are detected, whereas a T_d of four seconds means that many shorter events are missed. The relatively short overall optimum threshold compared to the minimum apnoea duration (two seconds versus five seconds) illustrates the fact that the duration of a flat region does not reflect the duration of an associated breathing pause.

When using the algorithm, clinicians can adjust the parameter settings to alter the performance for a particular recording. Typically, the algorithm would be run and the clinician then views the results. Based on their experience, they may decide to increase the sensitivity of the algorithm to ensure that all apnoeas are detected, especially with high amplitude recordings. On the other hand, there may be an excessive number of events detected, and if many of these appear to be false positives, then the sensitivity could be decreased. Increasing the standard deviation threshold D, decreasing the window L, and decreasing the minimum duration D all increase the sensitivity of the algorithm, but decrease the specificity (and vice-versa).

The performance of the system using the optimum settings is shown in Table 5.3; the results are relative to B_1. The overall high false positive rate (69%), and the low false negative rate (4.5%) reflect the importance of detecting apnoeas versus avoiding false detections. For individual recordings, there are variations: patient 2 has a 17% false negative rate, but three patients have a 0% false negative rate. The high rate for patient 2 is likely to be due to the fact that the recording is of high amplitude, with the signal saturating frequently (see Section 2.3.3); as mentioned above, a D of 5.0 was the optimum for analysing the patient 2 recording.

Patient	All apnoeas	Computer Detection		
	(human expert)	Total Detections	False +ve %	False -ve %
1	108	227	53	2
2	72	71	15	17
3	145	269	48	4
4	97	257	62	0
5	39	374	90	0
6	36	514	93	3
7	47	79	47	11
8	14	47	70	0
9	44	92	53	2
10	17	70	76	6
Average	-	-	**60 ± 20**	**5 ± 4**
All	**619**	**2000**	**69**	**4.5**

Table 5.3 Computer detector performance using the optimum settings of L = 1.1, D = 2.0 and T_d = 2. The false positive and false negative percentages are calculated as described in Section 4.2.2.

The difference in the length of apnoeas calculated using the algorithms in 5.1.2 compared to human experts is presented in Table 5.4. Of the reference apnoeas detected, most (77%) had their duration measured to within one second of the duration estimated by the experts. Whilst the duration calculation was not totally accurate, it was correct to the desired degree of accuracy in the majority of cases. The duration measurement was secondary to the actual detection of events because the experts placed much less importance on the duration accuracy compared to the actual

Patient	All apnoeas	2+ seconds less than experts	%	same ± 1 sec	%	2+ seconds longer than experts	%
1	106	*8*	8	*90*	85	*8*	8
2	64	*39*	61	*15*	23	*10*	16
3	139	*8*	6	*120*	86	*11*	8
4	97	*10*	10	*82*	85	*5*	5
5	35	*6*	17	*20*	57	*9*	26
6	40	*2*	5	*34*	85	*4*	10
7	14	-	0	*14*	100	-	0
8	39	*2*	5	*33*	85	*4*	10
9	42	*5*	12	*29*	69	*9*	21
10	17	*1*	6	*15*	88	*1*	6
Total	**584**	***81***	**14**	***452***	77	***61***	**10**

Table 5.4 Comparison of lengths of apnoeas as measured by computer and human expert.

detection of events. Thus the overall results were adequate in the context of this system. The only unusual result was for patient 2, where 39 out of 64 detected events had their duration underestimated by at least two seconds. The reason was likely to be due to the high amplitude nature of the recording, as noted previously, and hence the amplified cardiac signals or other noise causing spurious peaks to be detected, or causing variations in the flat regions of apnoeas such that they are no longer detected as flat regions. Fortunately, although patient 2 was one of ten recordings in B_1, there have been very few other recordings that have been of a similar nature; over the last three years, almost all recordings have been of lower amplitude, and therefore the discrepancy with patient 2 is not considered to be a problem in relation to the performance of the system as a whole.

The system has been in clinical use for several years, and has proven itself as a useful tool for clinicians [Ford, et al. 1992, Macey, et al. 1995, Tappin, et al. 1996a]. Although the system has a significant false positive rate, it reliably detects the majority of apnoeas, so that an expert can then view the detected events and determine which are apnoeas. However, it has not been used as a stand-alone apnoea detection system, and is always used in conjunction with a human expert. The system is not suitable as a scoring system due to the high false positive rate. Even for large studies, where individual apnoeas are of no concern, expert input is needed [Tappin, et al. 1996b, Tappin, et al. 1997]; an example of the system in use is presented in the following section.

5.2 Analysis of Data from a Study of Normal Infants

This section describes the analysis of a large volume of physiological data collected during a study of normal infants in the home. Analysis methods and some results are presented, and in particular, the apnoea detection system presented in Section 5.1 was used to detect apnoeas in conjunction with a human expert. The aim of this section is to demonstrate an application of apnoea detection, and as a part of that, develop more general analyses of physiological data in order to describe a variety of physiological patterns.

A recent study of normal infants collected a large volume of HomeLog data that consisted of over 400 nights of breathing and temperature recordings [Tappin, et al. 1996a]. Methods were designed to analyse this particular data set. Although the results of the research are primarily relevant in a medical context, the techniques used to calculate the results and the reasoning behind them are important, because they objectively define how the results were calculated and allow the methods to be reproduced elsewhere.

Given that one night's recording of breathing alone might contain 500,000 or more sample values (14 hours recorded at 10Hz [Ford, et al. 1992]), and there are over 400 nights of recordings, the data needs to be reduced to a manageable size in order to get meaningful results. For each night's recording, the data was therefore reduced to a number of measures that described the behaviour being studied. This approach has been used in other studies [Hoppenbrouwers et al. 1978, Schechtman et al. 1990]. Comparisons could then be made between different nights and different infants in relation to these measures.

Physiological results are briefly described [Ford, et al. 1996, Tappin, et al. 1996a, Tappin, et al. 1996b, Tappin, et al. 1997]. From the general physiological results and also the results in terms of apnoeas, the analyses can be evaluated in terms of their effectiveness at describing physiological patterns or behaviours.

5.2.1 Normal Infant Study: Analysis Requirements

The main aim of the study was to gather information on control infants which could be compared with information gathered from unwell or high SIDS risk infants. This study was to address the problem that there were many studies that recorded physical functions of high SIDS risk infants [Brown, et al. 1990], but few that recorded physical functions of normal infants, and hence there was little reference of what was considered normal behaviour. A secondary aim was to discover information regarding the physiological effects of vaccinations, and therefore the study was performed over vaccination periods.

Polysomnographic studies were performed in infants' homes, recording breathing and temperature data [Tappin, et al. 1996a, Tappin, et al. 1997]. Twenty-one infants were studied, each of whom had a low risk of SIDS: they were not a first child, they were breast fed, they slept on their side or back, they had no SIDS siblings, they had not experienced an ALTE, their parents were non-smoking, and their parents were older and not separated. The babies were studied in their homes using HomeLog [Ford, et al. 1992]. The recordings were performed around the times of vaccinations, and ideally each infant was studied for seven nights around the vaccinations at six weeks, three months and five months, with an additional three trial nights about the age of two weeks. The signals recorded were Graseby breathing, and five temperatures: rectal, anal, shin, abdomen and environment. The data took approximately one year to collect, corresponding to over 400 nights of recordings. The full methodology is published elsewhere [Tappin, et al. 1996a].

The data needed to be analysed to produce useful information or results. Firstly, given the number of recordings, the data needed to be reduced to a manageable size. The aim of any analyses was essentially to evaluate nightly patterns so that different nights could be compared, whether they were recorded from the same infant, or from different infants. Physiological patterns and behaviours relating to a night's sleep needed to be defined, and measures of these patterns

and behaviours could then be developed. Analyses were needed to calculate the measures from the raw data, and having calculated a number of measures for each night, comparisons could be made using statistical analyses of these measures.

Some of the behaviours of interest for the breathing and temperature signals are described. For breathing, the *breathing rate* and *breathing rate variability* are useful measures that describe the state of the child [Hoppenbrouwers, et al. 1980a, Harper, et al. 1987, Tuffnell 1993]. The change in breathing over time relates to sleep state, varying from shallow, regular breathing (quiet sleep) to higher amplitude, irregular, active breathing (REM sleep). The breath rate variability tends to be low during shallow, regular breathing, and higher during active breathing. The actual breathing rate is also of interest: breathing rate relates to temperature, and hence a hypothesis to be tested is that if a vaccination increases a baby's temperature, then the baby's breathing rate should also increase after a vaccination. The state of breathing tends to be consistent over a period of many minutes, typically about half an hour, corresponding to the time in REM or quiet sleep; this can be seen in Figure 2.12. thus breathing rate and breathing rate variability are two measures that describe breathing.

In terms of temperature, it has been shown that infants' temperatures oscillate during a night's sleep, and the amplitudes and periods of the oscillations are of interest [Brown, et al. 1992]. In particular, the basic information desired was whether the temperature was rising or falling, and this could be given by the rate of change of temperature. Each cycle appeared to have a duration of around one hour so the changes are not rapid [Brown, et al. 1992, Griggs, et al. 1995]. The average temperature of the infant is also of interest—the temperature would be expected to rise the night after the vaccination, as infections lead to raised temperatures, and a vaccination is a mild infection. Thus, two simple measures that describe the temperature behaviour are temperature itself, and the rate of change of temperature.

Apnoeas are an additional description of a breathing behaviour. It is not known whether apnoeas are associated with particular patterns of breathing or temperature, and therefore to discover any physiological patterns associated with apnoeas, they need to be detected. The physical behaviour at times of apnoeas can then be compared with the physical behaviour at other times. Thus, the time and duration of apnoeas are physiological measures that are calculated.

The analyses are to reduce the raw data from the recordings to measures of breathing rate, breathing rate variability, temperature, rate of change of temperature, and apnoea times and durations. These measures can then be further evaluated using statistical techniques. However, this section is concerned with reducing the data to measures of features, and the statistical analyses are described elsewhere [Tappin, et al. 1996a, Tappin, et al. 1997].

5.2.2 Calculating Physiological Measures from Raw Data

Specific analyses and methodologies were developed to measure the physiological behaviour described in the previous section. The analyses involved calculating measures of physical functions from the raw recorded data. The physical functions to be measured are described in the previous section, and this section describes the algorithms and methodology used. The statistical analyses are only briefly outlined in the following section, as they have been published elsewhere [Tappin, et al. 1996a, Tappin, et al. 1997].

Measures relating to breathing and temperature patterns as described in Section 5.2.1 were calculated for each recording, and these measures related to periods of the recording, as opposed to an entire night's recording. Therefore, the first processing that was performed was to segment the data into sections, labeled *epochs*. Epochs have been used by other groups to measure physiological variables [Scholten and Vos 1981, Harper, et al. 1987]. Each epoch has a start time and duration, and corresponds to sections of the signals that were recorded during that time. An epoch E can be defined from a start time t_{es} and an epoch duration d_e as:

$$E[t_{es}, t_{es} + d_e] = \{B, T_{rect}, T_{anal}, T_{shin}, T_{abdo}, T_{env}\} \tag{5.21}$$

where the sets B, T_{rect}, T_{anal}, T_{shin}, T_{abdo}, and T_{env} correspond to segments of the breathing signal $s(t)$ and the temperature signals $c_{rect}(t)$, $c_{anal}(t)$, $c_{shin}(t)$, $c_{abdo}(t)$ and $c_{env}(t)$ (corresponding respectively to rectal, anal, shin, abdominal and environmental temperatures) during the time of the epoch:

$$\begin{aligned}
B &= \{s(t): \quad t_{es} \le t \le t_{es} + d_e\} \\
T_{rect} &= \{c_{rect}(t): \quad t_{es} \le t \le t_{es} + d_e\} \\
T_{anal} &= \{c_{anal}(t): \quad t_{es} \le t \le t_{es} + d_e\} \\
T_{shin} &= \{c_{shin}(t): \quad t_{es} \le t \le t_{es} + d_e\} \\
T_{abdo} &= \{c_{abdo}(t): \quad t_{es} \le t \le t_{es} + d_e\} \\
T_{env} &= \{c_{env}(t): \quad t_{es} \le t \le t_{es} + d_e\}
\end{aligned} \tag{5.22}$$

Each recording included varying physical states, and therefore the physical behaviour of the infant was different during different epochs: awake or asleep, REM sleep or quiet sleep, beginning or end of the night, feeding, temperature oscillating, temperature rising or falling, or an apnoea occurring. Initially, epochs were defined according to the physical state at the time, but this caused problems in defining exactly what regions of the signal had a particular physical state, and in redefining the epochs whenever there was a change in the definition of a particular physical state. Therefore, the recordings were divided into epochs regardless of the physical state by taking adjacent epochs, starting at the beginning of each recording. The epochs were later classified according to the physical state at the time of their occurrence.

The duration of the epochs d_e was a parameter of the analysis, and all epochs were of the same duration. The actual value of the duration d_e was preset based on the physical behaviour being measured. The breathing rate and temperature change over a period of time in the order of minutes, as opposed to seconds. Temperature oscillations have been reported to be in the order of one hour [Brown, et al. 1992, Griggs, et al. 1995], and hence to measure rising or falling regions, the epoch length needed to be significantly less than 30 minutes. Previously, a one minute epoch had been used for calculating sleep state [Harper, et al. 1987], but that was only measuring breathing variables, not temperature which changes at a slower rate. Hence a five minute epoch was used, as it was considered long enough that changes in temperature would be clearly represented and short enough that variations in breathing would still be distinguishable. Other groups have typically used one minute epochs for breathing and heart rate variables [Hoppenbrouwers, et al. 1978, Scholten et al. 1985], although Scholten at al. [1981] have also used a three minute epoch. Using a five minute epoch and calculating the variables required, the volume

of data was reduced by a factor greater than 200, which for the total data recorded during the study corresponded to a reduction from 1.5GB to less than 7.5MB.

Having defined the epochs, the physiological measures were calculated. The breathing rate was calculated as the inverse of the median breath length, and the breath lengths during the epoch were calculated as the length between breaths in the breathing signal. A peak-to-peak detection algorithm was used to detect breaths [Tuffnell 1993]. This particular algorithm was selected because it had been successfully used with the Graseby breathing signal [Tuffnell 1993]. One problem with the peak detection algorithm was that it needed to detect several peaks and troughs before the parameters of the algorithm adapted to the amplitude and nature of the signal, and the algorithm was therefore considered inaccurate for the first ten to 30 seconds of an epoch, or when there was a large change in amplitude. The result was that there may have been spurious breaths detected or breaths missed [Tuffnell 1993]. Therefore, the median breath length was taken as opposed to the mean, as the median is less susceptible to noise. Similarly, to get a measure of the variability, the interquartile range of the breathing rate was calculated from the upper and lower quartiles of breath lengths (75^{th} and 25^{th} percentiles respectively). Both the median and the interquartile range are robust operators [David 1981], and reduced the influence of any errors in the peak detection algorithm.

Next, the temperature variables were recorded, which were more direct measures than the breathing variables. The mean value of temperature over the epoch was used as a measure of the absolute temperature, and the mean of the difference between adjacent samples was used as the measure of the rate of change of temperature. These variables were calculated for all temperatures.

The next step was checking the epochs to ensure that they contained valid data. There were occasional times when instruments had been switched off or sensors damaged, and the recorded data was of no use. Because one of the aims of the study was to compare different physiological patterns, the analyses were performed using epochs that include all signals. If no breathing was recorded (only temperature) then the data during that epoch was not considered useful, even if the temperature signals appeared accurate. Errors in the recorded breathing signal occurred when the instrument was turned off, and there would be a gap in the recorded data. The other common error was that the sensor would be disconnected, in which case the signal appeared as a flat line below the minimum normal value (below -100.0 in the case of the Graseby). If either of these errors occurred, it was usually when the child was awake and being handled by parents, and the data would not have been useful anyway. Thus, any epoch with missing breathing data or containing breathing data that included sample values below -100.0 was not considered for analysis.

Another problem with the recorded data was that there were periods during the recordings that corresponded to the child being awake. There were necessarily times at the start and end of each recording when the infant was awake, as the probes were connected before the infant was put to sleep, and the probes were removed once the infant woke up. Babies also woke up for feeding, especially the young ones, and so in several recordings there were periods of around ten to twenty minutes that corresponded to the infant being awake. Occasionally, an infant had trouble sleeping during the night and remained unsettled for up to an hour. The aim of this study

was to investigate behaviour during sleep, and therefore any epochs that occurred during periods where an infant was obviously awake were excluded. However, the method of determining whether the infant was awake or not was not precise.

Given a recording of breathing, there are two methods of determining periods during which an infant was awake: firstly, a human expert can interpret the data, and secondly, an algorithm that calculates sleep state can be used [Harper, et al. 1987]. The latter relies on the recording consisting of a certain approximate percentage (10%) of the data recorded while the infant was awake, and therefore requires some expert interpretation before the analysis can be performed. As an initial approach to excluding some of the awake data, the experts marked the times when the infant was unmistakably awake. Experts marked the start and end of each night's recording, and they also marked other times during recordings when the baby was obviously awake. Epochs that occurred at or overlapped these times were not considered for analysis.

There were a variety of physiological states represented within each recording, and as much as possible, each epoch was associated with the physical behaviour of the infant at the time, as far as it could be determined. Sleep stage was another physiological variable of interest. It is possible to calculate sleep stage using an algorithm based on the median breathing rate and breathing rate variability [Harper, et al. 1987]. Because the median breathing rate and breathing rate variability were calculated for all epochs, the sleep stage algorithm was used in the statistical section. Another method of determining the sleep state was to view the Graseby breathing signal with an entire night's data on one graph or screen, and mark out the regions of REM and quiet sleep by marking the regions of regular and active breathing. As explained in Section 2.3.3, displaying the Graseby signal over a period of hours allowed periods of active and regular breathing to be distinguished [Tappin, et al. 1996a]. Therefore, the experts studied all recordings and as well as marking the times that corresponded to the infant being awake, they marked the times when the breathing was active or regular.

The apnoea detection was performed independently of the epoch analyses. All apnoeas at least five seconds long were detected, and their duration measured. The detection was performed using the detection algorithm in Section 4.4 as an aid to a clinician. Thus, the majority of apnoeas were detected, and each event that was detected was checked by a human expert, both for its duration and whether it represented a pause in breathing or not. The apnoeas were matched with epochs in the statistical section of the analyses, and each epoch had an apnoea density calculated in terms of numbers of apnoeas per hour.

The result from applying the combinations of analyses to the recorded data was that the data were reduced to a number of epochs, each of which included measures of median breathing rate, breathing rate variability, temperature, and temperature rate of change. Each epoch was associated with the following physiological descriptions:

- age of infant;
- nights before or after vaccination;
- expert breathing state (regular, awake or other).

The apnoeas and their durations were recorded separately. All these measures were then processed during the statistical stage of the overall analysis to determine what patterns were present.

5.2.3 Physiological Results

This section presents some of the results obtained using the techniques described in the previous section. The statistical methods of combining the physiological measures are not described, but some of the physiological findings are presented. The aim is to illustrate how the analyses, including apnoea detection, lead to findings that are relevant in a medical context.

Overall, there were a total of 359 nights that were considered by the experts to have valid data, consisting of 156 nights prior to vaccination, 53 nights on the day of vaccination, and 150 nights following vaccination. These nights represented 3,609 hours of recordings, with an average of ten hours sleep per night. A total of 43,308 five minute epochs were defined, and the breathing and temperature measures described in Section 5.2.2 were calculated for each of the epochs. These measures were then passed to a statistical analysis programme, which is described elsewhere [Tappin, et al. 1996a, Tappin, et al. 1997].

Considering the various temperatures, the rectal temperature is the most consistent in terms of measuring core body temperature [Brown, et al. 1990], and hence the effect of vaccination on rectal temperature was considered. Compared to the rectal temperature during the three nights prior to vaccination, referred to as control nights, a significant increase in rectal temperature was seen the night after 48 out of 53 vaccinations. The mean increase was 0.58°C (±0.02), which is in

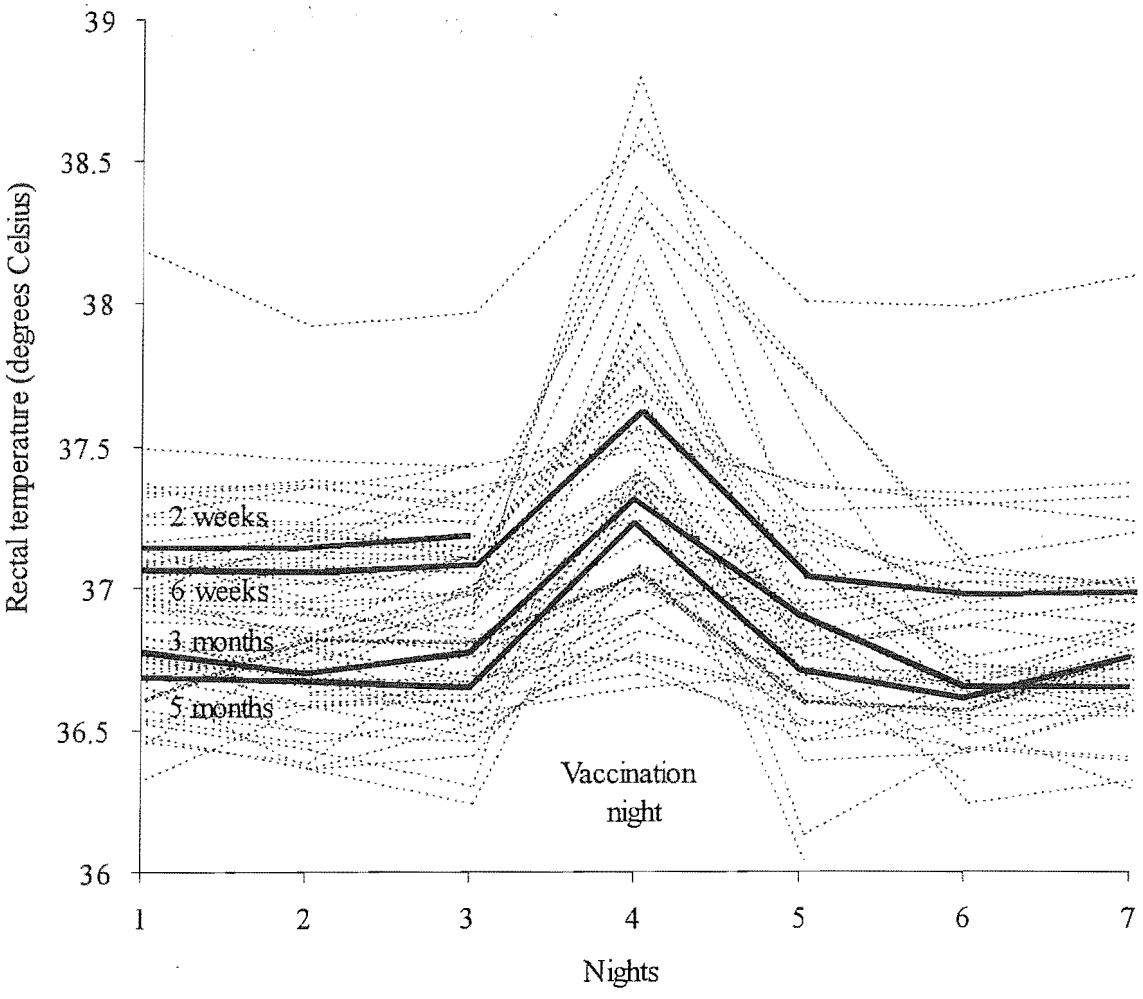

Figure 5.5 Mean rectal temperatures for each night of recording, the 4th night being after vaccination and illustrating the higher temperature. The mean of infants' temperatures is shown for the ages of 2 weeks, 6 weeks, 3 months and 5 months.

accord with previously published results [Rawson et al. 1990]. These results are illustrated by plotting the mean rectal temperature for each epoch over the usual seven nights of recording (three prior to vaccination and four after), as shown in Figure 5.5. The mean temperature usually returned to the pre-vaccination level by the second night after vaccination. By the third and fourth nights after vaccination, the mean temperature was the same as during the three control nights. There was a fall in the mean rectal temperature as the age of the infant increased, as illustrated in Figure 5.5.

The breathing rate also increased on the night after the vaccination, with a median increase of 9.5% (6.5, 12.5) relative to the control nights.

The results also show that breathing state is related to sleep state, a result that has been shown previously [Harper, et al. 1987]. The sleep state was calculated using an automatic classification algorithm (Harper sleep state [Harper, et al. 1987]), and the breathing was classified as regular or non-regular breathing by a human expert [Tappin, et al. 1996a]. Only control nights (nights prior to vaccination) were considered, and data at either end of each night's recording were excluded, leaving 13,646 epochs to be analysed. The Harper sleep state algorithm successfully classified 83.7% (11,420) of these epochs, with 16.3% (2,226) "undetermined," 1.8% (250) "awake," 55.7% (7,602) "REM," and 26.2% (3,568) "quiet." Regular breathing was present in 28.9% of the epochs, a similar percentage to the 26.2% of epochs classified as quiet sleep. Regular breathing had 70% sensitivity and 93% specificity when used to describe quiet sleep state, confirming that regular and active breathing relate to quiet and REM sleep.

The significance of this result was that other behaviours that were observed to be associated with active or regular breathing were likely to also be associated with sleep state. Hence, these relationships were studied. For example, the oscillations observed in infants' rectal temperatures were of interest [Brown, et al. 1992], as there appeared to be a relationship between rising temperature and active breathing, and between falling temperature and regular breathing. Therefore, a relationship between the oscillations and sleep state was expected: The relationship between rising temperature and REM sleep, and falling temperature and quiet sleep was confirmed by the results in the Table 5.5, which show that the temperature is more likely to be falling during quiet sleep and more likely to be rising during REM sleep.

Sleep State	**Rectal Temperature Changes**		**(%)**
(13,646 epochs)	**fall**	**static**	**rise**
quiet (3,568 epochs)	66	14	20
REM (7,602 epochs)	35	11	54

Table 5.5 Percentage of epochs with temperature changes in the "quiet" and REM Harper sleep state categories.

Apnoea detection was performed using the algorithm described in Section 4.4. The algorithm detected almost 90,000 possible apnoeas, which were then checked by one expert. The expert found that 28,322 of the 89,681 events were true apnoeas, resulting in a false positive rate of 67.6% (Table 5.6). Because the expert did not analyse the breathing signal directly, the false

negative rate cannot be calculated. The 359 nights of breathing signal and 28,322 apnoeas form a second reference set of apnoea as detected by human expert, and this set is referred to as B_2, as opposed to B_1 in Section 3.2.3. Note that B_2 is not as accurate as B_1 in terms of a reference, because B_2 represents the opinion of one expert only, and not all breathing signals have been analysed by that expert.

	Expert	Algorithm
Apnoeas	28,322	89,681
False positives f_p	-	67.6%

Table 5.6 Apnoeas detected from the home recordings of low-risk babies by expert and by algorithm. The expert apnoeas were detected by one expert checking all apnoeas found by the algorithm.

There were no apnoeas greater than 16 seconds in duration detected throughout the 359 nights, as seen in Figure 5.6. The lack of apnoeas greater than 16 seconds confirms that short central apnoeas are normal behaviour, and that longer central apnoeas are abnormal, and usually only associated with unwell or high SIDS risk infants. This finding is consistent with other studies [Hoppenbrouwers, et al. 1977]. Note that in the ten nights' recordings that constitute B_1, which are recordings from ALTE babies, two apnoeas of 22 seconds were recorded. For the normal infants, the number of apnoeas per night varied from seven to around 200, but every infant always experienced some apnoeas during a night's sleep. The distribution of apnoea durations and the frequency of apnoeas per night are shown in Figure 5.6, illustrating that the majority of apnoeas were less than ten seconds in duration.

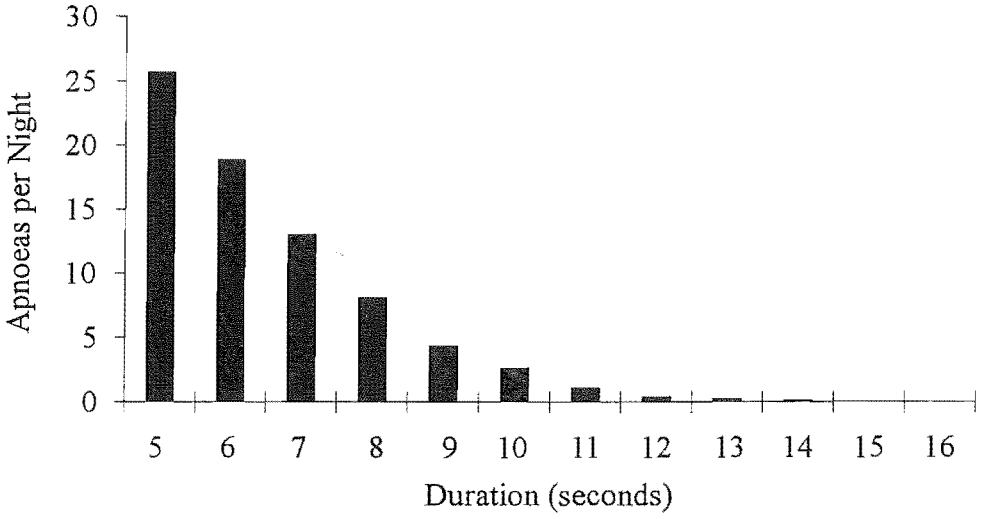

Figure 5.6 Frequency of apnoeas of each duration per night, where a night corresponds to nine hours of sleep. The frequency is the mean number of apnoeas per nine hours sleep for the control nights.

The vaccination affected the temperature and breathing, as mentioned above, and hence it was likely that the vaccination would also have affected the occurrence of apnoeas. By comparing the apnoea density on control nights with the apnoea density on the nights after the vaccination, it was found that there was a decrease in apnoea density on the night after vaccination (-29% [-20%,-37%] median reduction), as seen in Figure 5.7. However, if the minimum duration of the apnoeas was defined as three missed breaths (three times the median breath length, as calculated

from the median breathing rate), then there was no significant reduction in apnoea density on the night of vaccination (-5% [-33%,15%] median reduction).

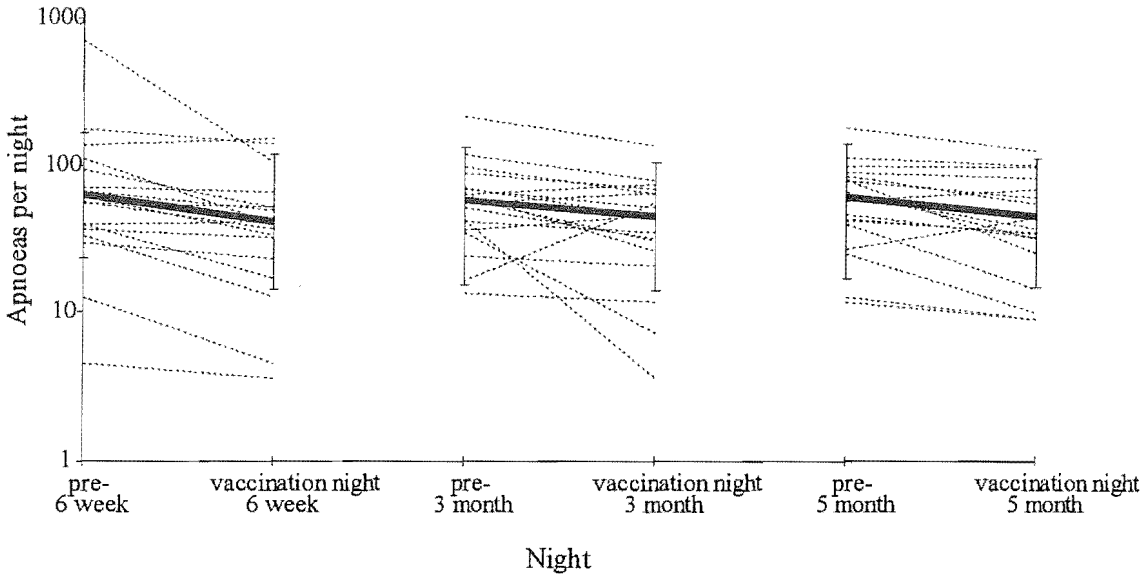

Figure 5.7 Apnoea density on control nights and nights after vaccinations, at six week, three month and six month vaccinations. One night is defined as nine hours of sleep, and the density refers to apnoeas five seconds and longer. The thick lines are the mean densities. The minimum duration of the apnoeas is five seconds.

However, if the minimum apnoea duration was set at three times the median breath length over the whole night, then the number of apnoeas remained the same for all nights. An explanation for this result was that during the night following vaccination, the respiratory rate was higher, corresponding to shorter breaths and leading to a shorter median breath length. Thus the minimum duration of three times the median breath length on the night after vaccination was shorter than during the control nights, and more apnoeas were included. This increase in the number of apnoeas during the night after vaccination meant that the numbers of apnoeas detected remained the same across both control and vaccination nights.

5.3 Discussion and Conclusions

The statistical method of apnoea detection has been used to detect possible apnoeas from a large number of studies, and greatly reduced the workload for clinicians. As part of a larger study of infant breathing and temperature, the apnoea detection algorithm and other physiological analyses have enabled physiological patterns in infants to be discovered and measured. The study presented in this chapter demonstrates that mathematical, objectively defined analysis methods can be used to analyse infant physiology.

Some analyses allowed physiological patterns to be measured and distinguished, while others did not. An example of the latter was that there was no correlation between apnoea density and other factors such as sleep state, temperature rising or falling, or breathing state. As well as reinforcing previous research regarding the physiology of normal infants, this research has led to new discoveries regarding temperature, sleep state and apnoeas.

There is certain to be further information to be discovered from the data, and it is also possible that the results described above could be described in further detail. However, as the purpose of the study was to detect and measure patterns of physiological behaviour, the analysis methods used were considered successful. While the analyses in Section 5.2.2 were not the definitive means of analysing the recorded data, they were a useful starting point.

One aspect of the analysis that may have led to inaccuracies was the five minute epoch length. The potential problem is that five minutes is a significant proportion of the typical time spent in REM or quiet sleep, which is in the order of half of the duration of a rectal temperature oscillation, or half an hour. Many epochs are likely to be defined during periods that include both REM and quiet sleep. Therefore, some epochs that were characterised as REM would include some quiet sleep, and vice-versa. The distinction between REM and quiet sleep state would be blurred to a certain extent, and any results would also be blurred. The extent to which the distinction was blurred is not known, but could be tested by repeating the analyses using a shorter epoch length and seeing whether the new results had increased distinctions between REM and quiet sleep state patterns.

Another feature of the overall analysis was that human expert involvement was required, and hence the results were not totally objective, even though the experts used objective guidelines. Expert involvement was required in determining the times during recordings that infants were awake, which was significant as the data during these times was excluded from any analysis. Marking regions of regular and non-regular breathing was also performed by experts, and experts also checked the events detected by the apnoea detection algorithm. The use of human experts for analysing the data means that the results may not be totally accurate, and may not be able to be reproduced exactly.

The analysis of the data recorded from normal infants allowed various physiological behaviours to be described. It is possible that further analyses may reveal other information. However, as a first analysis of a large volume of data, the method has allowed new information to be discovered and published, and provides a reference of normal physiological behaviour in infants. In terms of research studies, this is a typical example of how apnoea detection is used.

An apnoea detection algorithm was presented, and its application as a part of the analysis of a study of normal infants described. The algorithm detected the majority of apnoeas, with a high false detection rate. A human expert was required to check all detected events and reject any non-apnoeas. The algorithm has been in clinical use for several years [Dove, et al. 1990], and is considered a useful tool by clinicians.

The detection algorithm measured statistical properties of the breathing signal to detect breathing pauses. A low standard deviation corresponds to flat regions in the signal, and all signals that represent apnoeas contain a flat region. Thus, the flat regions were detected by detecting regions of low standard deviation, after which a second algorithm calculated the start and end points, and hence the duration. This process is similar to the manner in which experts detect apnoeas, as explained in Section 3.3: firstly they detect flat regions as possible apnoeas, and then evaluate these regions in more detail and classify them as apnoea or non-apnoea. However, having detected flat regions and checked that the associated pause is greater than the

minimum duration, the experts also check that the flat region does indeed correspond to a breathing pause, and it is this last check that the statistical method does not perform.

The results using the statistical method are consistent: the majority of apnoeas are detected, for a high rate of false detections. A previous method achieved a lower false positive rate (17%) but for a higher false negative rate (26%) [Bruckert, et al. 1982]. The performance of this previous method is less favorable than the statistical method presented, according to the performance measure described in Section 4.2. The method appears to be an effective means of detecting possible breathing pauses and measuring their duration, but does not discriminate between flat regions that represent pauses in breathing and flat regions that do not represent pauses in breathing. An expert can perform this test, but there can be a large number of events to check.

In conclusion, the statistical method is a useful aid to a clinician, but not a replacement. In order to achieve more accurate apnoea detection, further development is required. The development required appears to be in the area of determining exactly which flat regions in the signal correspond to definite breathing pauses. Any accurate apnoea detection algorithm must not only distinguish flat region in the breathing signal, but also distinguish between flat regions that represent pauses in breathing and those that do not. The following chapter addresses some of the problems with the statistical method of apnoea detection, developing both a more precise definition of apnoea and a more accurate detection algorithm.

Chapter 6

Expert System for Apnoea Detection

This chapter proposes solutions to some of the problems of apnoea detection. Firstly, a model of apnoea is developed based on signal properties that discriminate between apnoea and similar non-apnoea events. Secondly, a system is developed for classifying events using these properties. The whole system is labeled an expert system as it is performing the function of a human expert.

6.1 Introduction

Previous chapters have explained the problems of apnoea detection, and these problems can be grouped into three categories:

1. a lack of definitions of signal corresponding to apnoea;
2. inconsistent expert reference data;
3. a lack of accurate detection algorithms.

These problems are not independent of each other. The lack of definitions restricts the development of detection algorithms, as there are no definitive standards with which to measure performance. Similarly, inconsistent reference data means that definitions cannot be measured as exact. The ultimate solution to these problems must cover all three areas. Hence, the research presented in this chapter aims to be a step towards a final solution.

In this chapter, a new signal definition of apnoea is developed. The definition is based on expert interpretation. In particular, because human experts view the signal in a graphical form when detecting apnoeas, the definition is based on deterministic, or shape, properties of the signal. The properties are intended to be general properties that describe most apnoeas within a breathing signal, and that can be applied to other signals. The properties can only be developed based on existing reference data, which are not entirely consistent, and therefore the properties are unlikely to discriminate between *all* apnoea and non-apnoea events.

The second part of this chapter presents a new detection system that is based on the deterministic signal properties discussed above. These properties are a major component of the system, but the properties only give measures of characteristics at locations in the recording. Selecting locations and classifying them based on the property measures are additional functions for which the system is developed. When combined with properties, it is described as an expert system. The system models human expert detection in that the signal is quickly scanned for apnoea-like regions, and then these regions are studied in more detail before classifying them as apnoea or non-apnoea.

The results from the training and testing of the system allow both the properties and the system to be evaluated. The aim is not to perform an exhaustive analysis of the results, but rather to determine whether the properties and system are effective or not. Human expert opinion is the reference standard of what signals represent apnoeas, and because human expert opinion is inconsistent, there is little point in evaluating the system performance in detail based on individual

events that are missed or detected. The information that is valuable is a measure of how successfully the properties discriminate between apnoea and non-apnoea events, and of the detection accuracy of the system based on the properties. The result is a set of properties that discriminate between most apnoea and non-apnoea events, and an expert system that detects apnoeas more effectively than previous methods.

6.2 Model of Apnoea in an Abdominal Breathing Signal

This section presents properties of breathing signals that represent apnoeas. The properties are signal characteristics that discriminate between apnoea and non-apnoea events, and they are described mathematically based on signal characteristics used by experts.

6.2.1 Objectives

The aim is to develop a general set of properties that describe most apnoeas, with common features described systematically and mathematically. Ideally, the properties are general, meaning that they are independent of the particular instrument used, the baby, the sensor placement, and other factors. The properties measure similar information to that which an expert considers when viewing a breathing signal.

A deterministic property refers to a geometric or shape property of a signal. Shape properties appear to be well suited to describing apnoea as experts use the shape properties to recognise apnoeas—they do not look at the signal and transform it to the frequency domain, or calculate the statistical properties of the signal. There are two reasons for developing properties:

1. to reduce the raw data by removing redundant information;
2. to describe an apnoea signal mathematically.

Classifying from a small number of measures is an advantage, as long as those measures contain information that allows discrimination. Typically, breathing signals contain a large number of samples that do not directly relate to the likelihood of an apnoea having occurred. Ideally therefore, the properties are few in number and each has a high discriminating power. It is also important that each property measures different information, in other words that the properties are relatively independent. There is no advantage in having two similar properties. By the same token, it is also important that as much information as possible is measured by the properties, and that there are enough properties to measure information that allows apnoea and non-apnoea events to be distinguished. In the extreme, one property measure by itself would be of little value. Having developed a set of properties, classification is then performed by combining the properties in some manner, which is simpler than classifying directly from the raw data.

If the properties are general, then they can be applied to other signals. An apnoea in a breathing signal has common characteristics across all signal types—it is a flat region within surrounding breathing, as seen in Figure 2.4. Each signal type also has its own idiosyncrasies, and so properties developed using one signal type are likely to contain information that is only relevant to that signal. Hence, while the main aim is to develop properties that are as accurate as possible, it is also important that the properties describe general characteristics that apply to other signals, and represent the opinion of other experts.

If a group publishes findings relating to apnoea, as in Chapter 5, mathematical descriptions are a reference for other groups to either replicate work or evaluate findings with an objective definition of the term "apnoea." The definitions could also provide a starting point from which general apnoea detection standards could be developed.

The objective of the work in this section is therefore to develop properties of a breathing signal corresponding to apnoea, where each property consists of a description of the signal shape and a mathematical description giving a measure of the described shape. Ideally, the properties should:

1. discriminate between apnoea and non-apnoea events;
2. be small in number;
3. be independent;
4. apply to other signals;
5. represent the opinion of other experts.

6.2.2 Development of Properties

Properties are defined based on experts' descriptions. The procedure to describe the properties is shown in Figure 6.1. Starting with experts' observations, a common characteristic is extracted, described mathematically, and adjusted until it discriminates between apnoea and non-apnoea events. These steps are repeated for other characteristics.

6.2.2.1 Expert Interpretation and Properties in Signals

Human expert opinion is the reference standard for which signal corresponds to an apnoea, and therefore the first step is to gather the experts' opinion. Experts recorded their observations of the characteristics of the Graseby signal they used to distinguish between apnoea and non-apnoea events (step 1 in Figure 6.1). The result is a number of written descriptions relating to what signal shape characteristics are used to distinguish between apnoea and non-apnoea events.

The most general characteristics are developed into properties, step 2 in Figure 6.1. Similar observations from each expert are combined as one characteristic, forming the basis of a deterministic property, denoted x. Whether a shape characteristic is general or not is judged by the number of apnoeas in which that shape is observed. For example, all experts refer to flatness in one form or another so the first properties described relate to flatness. Starting with general descriptions, each property narrows the model of apnoea. A single property is not used to discriminate between all apnoea and non-apnoea events. Subsequent properties are designed to exclude specific non-apnoea events, whilst still including all apnoeas. Each property is described as in steps 2 to 7 in Figure 6.1, and then, if another property is required (step 8), the procedure is repeated.

B_1 is used as a reference set to test the discrimination (see Section 3.2.3). Having identified a property, say property x, a subset B_{1x} of B_1 is formed. B_{1x} contains all non-apnoea signals *without* property x, and B_{1x} also contains all apnoeas from B_1. B_{1x} is formed by viewing all events that are detected as being flat regions but that are not apnoeas, and selecting those that do *not* fit the description of x, step 3 in Figure 6.1. At this point consultation with the experts helps to clarify

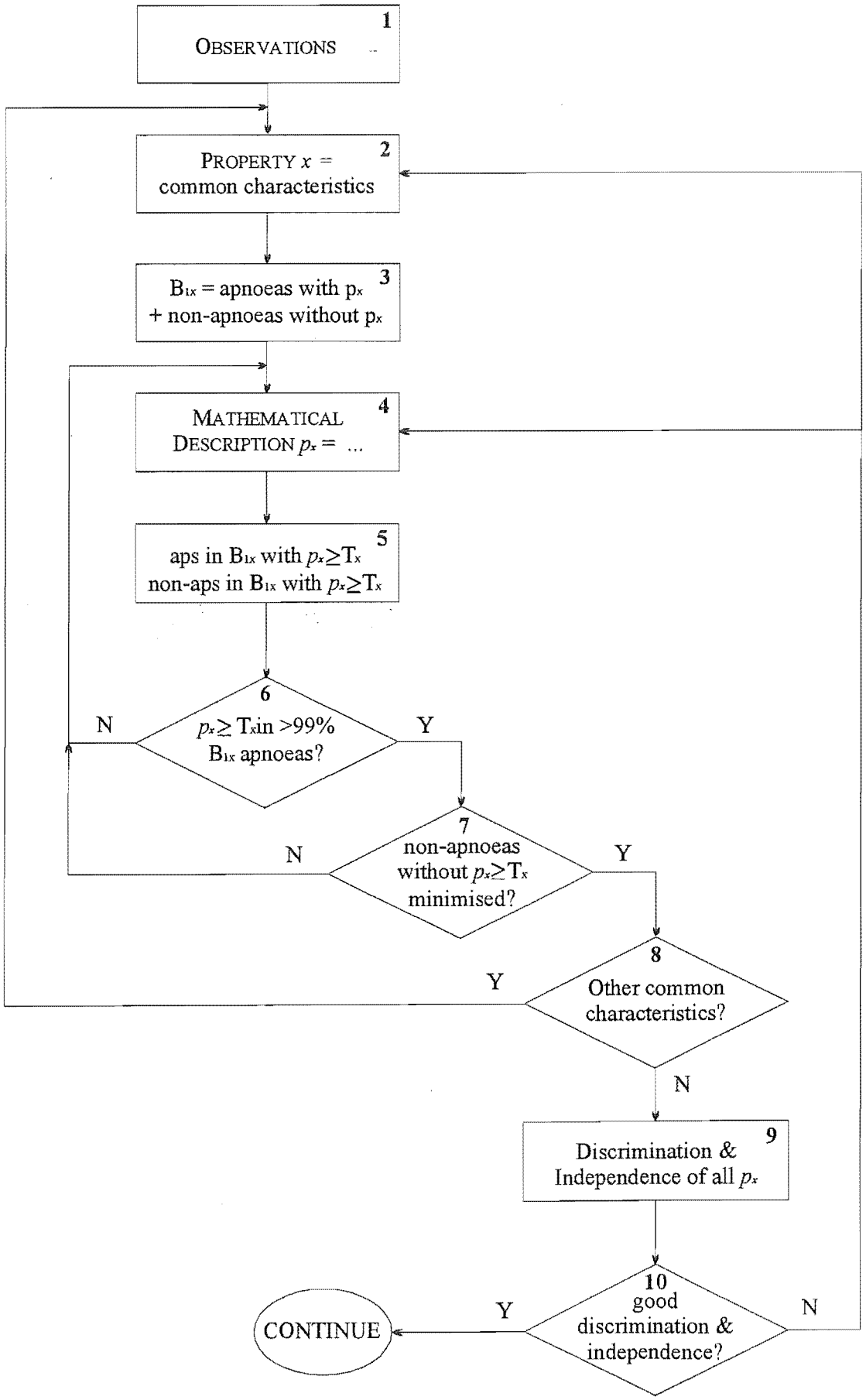

Figure 6.1 Procedure for developing property descriptions and definitions.

their observations. Therefore, all the non-apnoea events in B_{1x} can be said to not have the property x.

6.2.2.2 *Descriptions of Properties*

A property x is described mathematically to give a measure of x termed p_x (step 4 in Figure 6.1). Ideally, the function p_x is increasing with respect to apnoea likelihood and, in particular, extremely low values of p_x almost always correspond to non-apnoea events, and extremely high values of p_x almost always correspond to apnoeas.

The properties are studied to determine how well they distinguish between apnoeas and a set of similar non-apnoea events (steps 6, 7, 8 & 10 in Figure 6.1). The set of non-apnoea events is the set of false positives produced by a previous detection algorithm [Macey, et al. 1995]. A number of checks are performed, and the results are used as feedback for adjusting the properties.

The apnoea and non-apnoea events in B_{1x} are checked for the presence of property x as described by p_x. The measure p_x is calculated for all the B_{1x} signals, step 5 in Figure 6.1, and then x is defined as being present in a signal when p_x is above a threshold T_{px}, step 6 in Figure 6.1. The description is modified until T_{px} can be set such that almost all (> 99%) apnoea have the property ($p_x \geq T_{px}$).

Having set T_{px}, the number of non-apnoea events with $p_x \geq T_{px}$ is recorded. The description of p_x is modified to reduce the number of non-apnoea events with x, step 7 in Figure 6.1. By considering the characteristics of the non-apnoea events that have $p_x \geq T_{px}$, the property descriptions are adjusted with the aim of reducing p_x for non-apnoea events so that $p_x < T_{px}$, and at the same time ensuring that $p_x \geq T_{px}$ for at least 99% of all apnoeas.

The non-apnoea events remaining are then studied in step 8; if there are common shapes, a new property is developed (step 2). If there is no obvious common feature and there are few remaining non-apnoeas, then there may be sufficient properties. Deciding on an acceptable number of remaining non-apnoeas is based on two factors: 1. the overall desired detection performance figures (for example, 95% detection with a 15% false negative rate); and 2. whether any further common features can be identified.

6.2.2.3 *Parameter Tuning*

Having developed a set of properties, the mathematical descriptions are tuned. The properties are based on human interpretation, and humans are typically poor at differentiating between small differences, such as deciding whether a 1.1 or a 1.2 second window would be best for calculating a standard deviation. Each property measure p_x has tuning parameters added to its mathematical description, and by optimising these tuning parameters, such as the window length in the previous example, the discriminating power of each property is increased.

The optimisation is in terms of maximising the discrimination of the properties. A measure of discrimination is the power of discrimination (also called the power of test), which can be measured by the number of false detections for a given detection rate [Mood et al. 1974, Neter, et al. 1978].

However, although the percentage of false positives is a direct measure of performance, it is not necessarily a clear indicator of performance: small percentage changes in the false positive

rate can represent significant improvements in terms of an apnoea detection system, and vice-versa. For example, a 7% decrease in false positives from a rate less than 30% is more significant than a 7% improvement at higher rates: improving from 20% false positives to 13% false positives means changing from 1 in 5 events being a false detection to approximately 1 in 8 events being a false detection, a significant saving in terms of a clinician's time. On the other hand, improving from 80% to 73% false positives makes little difference to a clinician viewing the events. Thus, the false positive rate as a measure of performance does not relate directly to a human expert's experience.

As the performance evaluation is ultimately done by human experts, a measure is needed that relates more to the experience of the clinician than the false positive rate. The *Specificity Index* (SI) is defined as a function of the false negative rate, f_n (i.e. the percentage of missed apnoeas):

$$\mathrm{SI}(f_n) = \frac{N_{total}}{N_{false}}, \tag{6.1}$$

where N_{total} is the total number of events detected, and N_{false} is the number of false detections. SI is the number of apnoeas detected for every false detection; for example, SI(1) = 3 means that for 1% of all apnoeas not being detected there is one false detection for every three apnoeas detected. The Specificity Index is also a measure of the power of discrimination, in terms of discriminating between apnoea and non-apnoea events. The detection rate is typically in the order of 1% to 5%, relating to 99% and 95% confidence intervals [Mood, et al. 1974], so SI is maximised for false negative rates between 1% and 5%.

For the results presented, the parameters of each property measure are tuned to optimise SI(5), the case where 5% of all apnoeas are undetected. Each property measure is optimised individually because a global optimisation of all property measures is not computationally feasible. Each parameter is optimised across a wide range of values. If a parameter has discrete values, it is adjusted using the smallest possible increment; for example, the maximum duration would be incremented one sample at a time. If a parameter value is continuous, it is adjusted by increasingly smaller increments until no further significant improvements in performance are achieved. The result is a set of property measures with parameter values that optimise a measure of the power of discrimination.

6.2.2.4 Verification of Properties

A check of whether each property measure p_x discriminates between apnoea and similar non-apnoea events is performed.

The cumulative frequency curves of p_x values for apnoea and non-apnoea events are plotted. If the apnoea and non-apnoea curves for a particular property measure are different, then the property has discriminating power (step 9) [Mood, et al. 1974]. A cumulative frequency graph for a particular property has the proportion of events along the y axis, and the value of the property measure along the x axis. For the apnoeas, the proportion of events (y axis) that are of a value or less (x axis) are plotted, and the process is repeated for non-apnoea events. The curves are normalised individually along the y axis so that the y axis value represents the proportion of events of that value or less.

A test to measure the probability of two distributions being different is required; the test must be non-parametric as no assumptions are made about the distributions (as opposed to the Chi-square test, for example [Mood, et al. 1974]). A commonly used non-parametric test that is recognised in the international statistical community is the *two-sample Kolmogorov-Smirnov test*, also called the Smirnov test [Lindgren 1976]. The test is based on the maximum separation between the two curves, which is defined as the statistic D:

$$D = \sup_{\text{all } p} \left| F_m(p) - G_n(p) \right| \tag{6.2}$$

where "sup" is the supremum, or upper bound, and $F_m(p)$ is the first distribution function of m samples, and $G_n(p)$ the second distribution function of n samples [Lindgren 1976]; in this case F and G correspond to distributions of apnoea and non-apnoea events (or vice-versa) of a property measure p. The Smirnov test is a method of calculating a probability p that the distributions are the same (the null hypothesis), using the following equation:

$$\mathrm{p} = \exp\left(\sqrt{\frac{-2D^2}{\left(\frac{1}{m} + \frac{1}{n}\right)}} \right) \tag{6.3}$$

Hence, the larger the difference D and the greater the number of samples m and n, the lower p, and the greater the probability that the distributions are different, or in other words that the measure p discriminates between apnoea and non-apnoea events. If a property measure has a significant p value ($p > 0.001$) then the property is not used.

The properties are checked for independence to ensure that the minimum number of properties are used, as mentioned in the objectives described in Section 6.2.1. The independence of the p_x measures (step 9 in Figure 6.1) is determined by calculating correlation coefficients [Mood, et al. 1974]. The correlation coefficient between two properties is measured by correlating all values of the first property with all values of the second property, where the property values are measured for all apnoea and non-apnoea events in the reference set. A low correlation coefficient indicates that the two properties do not describe the same information. There is no exact figure below which the correlation is "low," but coefficients close to 1.0 mean that the properties are not independent, whereas coefficients close to 0.0 mean that the properties are independent. All properties are checked against each other to ensure that no two properties describe the same information.

6.2.3 Deterministic Properties of Apnoea

Following the procedure described in Section 6.2.2, a series of mathematical descriptions are developed forming a model of an apnoea signal. Four experts recorded the signal shapes, or characteristics, they used to distinguish apnoeas. According to their interpretations, an apnoea represented by a breathing signal is essentially a flat region. Therefore, the properties are developed as properties of flat regions. The characteristics are deterministic, meaning that they refer to geometric or shape features of the signal. The approach is firstly to define the flat regions, and secondly to define and measure properties of these regions.

6.2.3.1 *Flat Regions*

A flat region is defined as either present or not—there is no degree of being a flat region. The experts considered a signal to be flat relative to surrounding breathing so a signal is flat only in the context of breathing before or after the region. Given a sampled breathing signal $s(t)$, a region R is defined as a set of adjacent discrete samples:

$$R = \{s(t_i) : t_1 \le t_i \le t_2\} \tag{6.4}$$

where $s(t_i)$ is the sample value of the breathing signal at time t_i, t_1 is the start time of the region, and t_2 the end time. Standard deviation has been used as a measure of flatness in a previous system, enabling successful detection of 99% of apnoeas [Macey, et al. 1995], so a flat region R_f is defined as a region with low standard deviation:

$$R_f = \{s(t) : \max(\sigma[s(t_i)]_{L_f}) \le M_f \quad \forall\ s(t_i) \in R\} \tag{6.5}$$

where σ is the standard deviation calculated over a window L_f, and M_f is a standard deviation threshold. The start time of the flat region is labeled t_s and the end time t_e. A flat *region* is also defined to be flat for a period of time greater than a minimum duration T_r:

$$t_e - t_s \ge T_r \tag{6.6}$$

An example of a flat region is shown in Figure 6.2. Flat is a relative term and therefore M_f is varied with the surrounding breathing amplitude:

$$M_f = k_f A \tag{6.7}$$

where A is a measure of the surrounding breathing amplitude, and k_f is a tuning parameter. The aim in defining the flat region is to include all possible apnoeas. M_f is therefore set to the minimum required to detect almost all (> 99%) of the flat regions that occur within apnoeas.

When detecting the apnoeas in B_1, the experts viewed ten to 15 seconds of breathing prior to a possible apnoea event. The breathing viewed after a possible apnoea event was of short

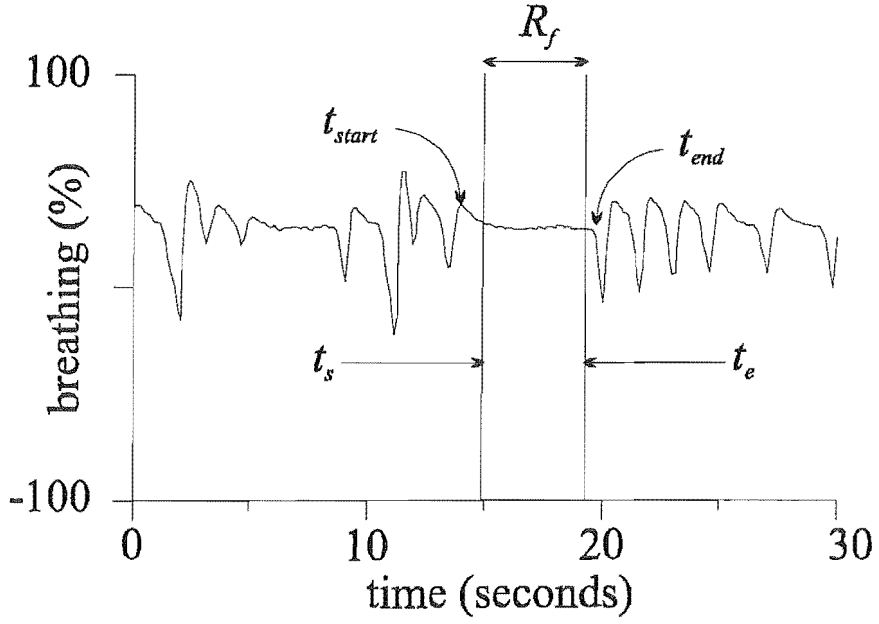

Figure 6.2 Apnoea with flat region R_f starting at t_s and ending at t_e; apnoea starts at t_{start} and ends at t_{end}. The vertical axis has a relative scale with the minimum output of the Graseby corresponding to -100%, and the maximum to 100% (see Section 2.3.3). A pause shorter than five seconds occurs prior to the apnoea.

duration, usually less than ten seconds and often less than five seconds. Therefore, the signal prior to a flat region is taken to be the region relative to which the region is flat, and this region of signal R_p is defined:

$$R_p = \{s(t_i): t_s - T_p \le t_i < t_s\} \tag{6.8}$$

where T_p is the duration of the region, and the flat region starts at time t_s. An example of R_p is shown in Figure 6.3 for T_p = 15 seconds.

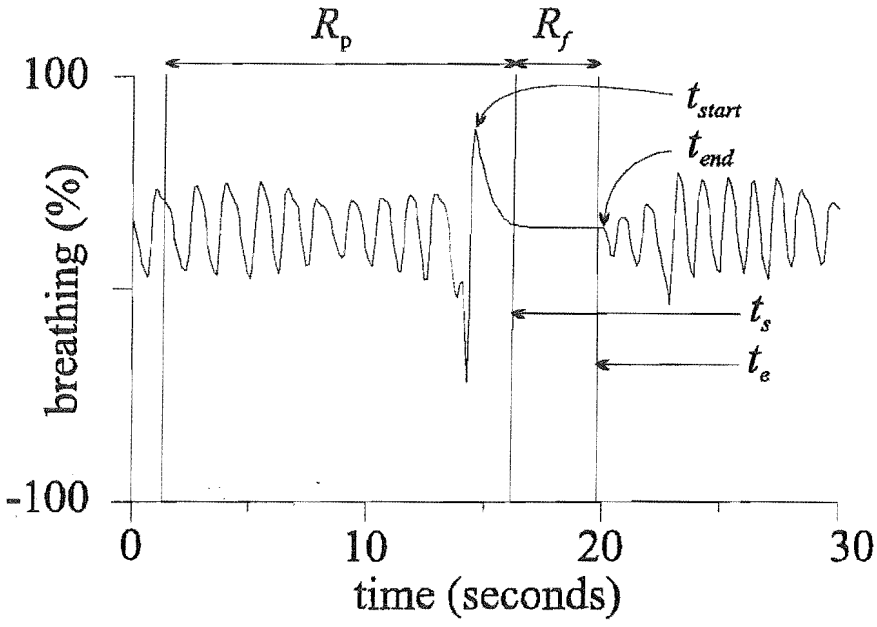

Figure 6.3 Apnoea preceded by sigh. Flat region R_f is preceded by a 15 second region R_p which includes a sigh that is represented by a single large amplitude oscillation just before R_f.

An indicator of the amplitude is the *range* of the amplitude of normal breaths prior to an apnoea. Thus, a measure of the range of the signal in R_p could give an estimate of amplitude A. However, a sigh often precedes an apnoea, and this is seen as a large peak and trough in the signal (see Figure 6.3). The minimum to maximum range of R_p in Figure 6.3 would be the range of the sigh, but not the range of the signal during normal breaths. The minimum to maximum range is therefore not a reliable measure for A, but if the extreme high and low values during the sigh are excluded from R_p, the range of the resulting subset is a more accurate measure. Another set r_p of signal values corresponding to normal breathing is therefore defined:

$$r_p = R_p \cap r_s \tag{6.9}$$

where r_s is the set of extreme points that occur during any sigh, and $r_s \subset R_p$. The range of r_p is the range of normal breathing, and is defined as V_p:

$$V_p = \max(r_p) - \min(r_p) \tag{6.10}$$

Thus, a measure of amplitude A is defined:

$$A = \begin{cases} 20\log_{10} V_p & V_p \ge 1.0 \\ 0 & V_p < 1.0 \end{cases} \tag{6.11}$$

Since A in (6.11) would not be defined for $V_p < 1.0$, A is set to zero in that case. Given that the signal has a possible range of 200%, an amplitude range of 1.0% or less is very small and

therefore $A = 0.0$ is an appropriate approximation. Hence, no apnoeas can be detected if the respiratory variation measured by the Graseby is less than or equal to 1% of its range.

The set r_s is the set of extreme values during any sighs in R_p, and can be approximated by the extreme values in R_p. The extreme values of R_p can be defined by *order-statistics*; order-statistics are any statistic that is derived from a ranked set of samples [David 1981]. The j^{th} order-statistic of R_p is defined as $s_{(j)}$ where $s_{(1)} \leq s_{(2)} \leq s_{(3)} \leq \ldots \; \forall \; s \in R_p$. Thus, the highest and lowest values of R_p can be removed by excluding $s_{(j)}$ where j is either close to 1 or close to N_p, and N_p is the number of elements in R_p. Hence, an approximation to V_p in (6.10) is defined by removing the highest and lowest y% of R_p:

$$V_p \approx s_{(j1)} - s_{(j2)} \tag{6.12}$$

where $j1 = \left[\frac{(100 - y)}{100} N_p \right]$ and $j2 = \left[\frac{y}{100} N_p \right]$, and $0\% \leq y < 50\%$. [.] is the integral function so that $j1$ and $j2$ are positive integers. Thus, for example, if $y = 0\%$, then the amplitude is calculated from the full range of R_p. The value for y was set by comparing A with the minimum M_f required to detect the flat regions of all apnoeas in B_1 (the reference set of ten recordings and 619 apnoeas as found by three experts—see Section 3.2.3), where A was calculated for values of y ranging from 0% to 49% in 1% steps. A for $y = 13\%$ has the most consistent relationship with the minimum M_f, where the relationship is considered consistent if it is increasing. Therefore, in order to calculate a measure of amplitude, 13% of the signal is excluded due to the presence of sighs and other behaviour.

From equations (6.7) to (6.11), the standard deviation threshold M_f is defined. The parameters T_r, k_f and L_f are tuned to minimise f_p for an f_n of 5%. A full grid optimisation of every combination of the three parameters, each across its entire practicable range, is performed using B_1. An apnoea is considered detected if the start of the flat region is before the end of the apnoea and the end of the flat region is after the start of the apnoea (see Section 4.2.1). The resulting values are $T_r = 2.8$ seconds, $k_f = 9.8$ and $L_f = 1.1$ seconds (see Table 6.1).

6.2.3.2 Properties of Flat Regions

Four deterministic properties of a flat region corresponding to apnoea in a Graseby signal are described:

1. *Flatness*: the degree to which the flat region of the apnoea is flat;
2. *Duration*: the duration of the event;
3. *Thinness*: the degree to which the flat region is thinner than surrounding breathing; and
4. *Smoothness*: the degree to which the flat region is smooth or regular.

6.2.3.2.1 Flatness

The first property of a flat region is *flatness*, which is a measure of how flat the flat region is relative to the overall signal. A measure of flatness p_1 can be split into two parts: an absolute measure of the flatness of the flat region, say y_f, and a measure of the amplitude of the surrounding breathing, which is already described as A in (6.11). The measure p_1 is defined:

$$p_1 = y_f A \tag{6.13}$$

Property	Performance	Parameter	Optimum	Minimum	Step	Maximum
Flat	P = 24.51	L_f	1.1s	0.7s	0.2s	1.9s
Regions	f_n = 2/619	k_f	9.8	9.0	0.1	12.0
	f_p = 1184	T_r	2.8s	1.6s	0.1s	3.1s
		T_p	17.5s	15.0s	0.5s	19.5s
p_1	SI(5) = 1.61	L_1	0.9s	0.3s	0.2s	4.5s
p_2	MSE = 0.38	T_e	0.3s	0.1s	0.1s	1.0s
		k_{max}	-0.89	-1.2	0.01	1.0
		k_{min}	0.02	-1.0	0.01	1.0
		T_{win}	2.3s	0.0s	0.2s	3.1s
		T_{pk}	0.8s	0.1s	0.1s	1.5s
		T_s	1.9s	0.1s	0.1s	4.0s
		T_{min}	3.8s	2.0s	0.1s	5.0s
p_3	SI(5) = 2.31	L_3	3.5s	0.3s	0.2s	4.5s
		k_s	1.05	0.0	0.02	1.2
		k_e	0.65	0.0	0.02	0.9
p_4	p_{41} SI(5) = 1.58	L_4	0.3s	0.1s	0.2s	1.5s
	p_{42} SI(5) = 1.46	$T_{shiftst}$	2.8s	0.0s	0.1s	4.0s
		$T_{shiftend}$	-0.2s	-1.0	0.1s	1.0s

Table 6.1 Optimised parameter values for property definitions, based on training with B_1. MSE refers to the Mean Square Error.

The greater p_1, the flatter the region and hence the more apnoea-like. As in (6.5), the absolute flatness y_f is defined by the mean of the standard deviation of R_f:

$$y_f = \frac{\left(\frac{\left(t_e - \frac{L_1}{2}\right) - \left(t_s + \frac{L_1}{2}\right)}{t_{sample}} \right)}{S_{sum}} \tag{6.14}$$

where

$$S_{sum} = \sum_{t=t_s+\frac{L_1}{2}}^{t=t_e-\frac{L_1}{2}} \sigma[s(t)]_{L_1} \tag{6.15}$$

where L_1 is the window length over which the standard deviation is calculated, t_{sample} is the time between samples, and $\frac{L_1}{2}$ is rounded so that it is a multiple of t_{sample}.

6.2.3.2.2 Duration

The second property of a flat region R_f is its *duration*. There are two durations to consider. The first is the duration of the flat region itself, and is defined by t_s and t_e as in (6.6). The second is the duration of the apnoea within which the flat region occurs. For the latter, duration is measured by defining start and end points t_{start} and t_{end} of an apnoea. Figures 6.2 and 6.3 illustrate the different start and end points of apnoeas and flat regions.

The experts describe the start of an apnoea as the first significant peak or trough prior to the flat region, the point where the last breathing movement can be identified (t_{start} in Figures 6.2 and

6.3) [Bruckert, et al. 1982, Macey, et al. 1995]. After the peak or trough, the signal returns to a rest value, which is the start of the flat region. The decay curve could represent either a gradual breathing movement or signal artifact generated by the instrument, or a mixture of both. The reason *significant* peaks or troughs are considered is that there are many small peaks and troughs in the signal due to cardiac oscillations or other artifact, as on the region R_{f2} in Figure 2.9.

Given a peak or trough prior to R_f, its significance is determined in two ways. Firstly, the peak or trough is the extreme value relative to surrounding points; the insignificantly small peaks and troughs are assumed to be due to cardiac oscillations and other noise [Marshall 1986]. The second constraint is that the amplitude of the peak or trough is outside the range of the signal samples in R_f —a peak or trough at a similar amplitude to the flat region is not significant. Hence, the start time is defined as:

$$t_{start} = \max\left\{T, U, \min\left\{t_s, t_x, t_y\right\}\right\} \tag{6.16}$$

where T is the set of points of greater amplitude than the flat region, defined as:

$$T = \left\{t_x : s(t_i) > X_{pk} \;\forall\, t_i : t_i - T_{pk} \le t_x \le t_i + T_{pk},\; t_s - T_{win} \le t_x \le t_s\right\} \tag{6.17}$$

and U is the set of points of lower amplitude than the flat region, defined as:

$$U = \left\{t_y : s(t_i) < X_{tr} \;\forall\, t_i : t_i - T_{pk} \le t_y \le t_i + T_{pk},\; t_s - T_{win} \le t_y \le t_s\right\} \tag{6.18}$$

where t_x are the peak times and t_y are the trough times within T_{win} of the start of R_f. If no significant peaks or troughs exist (no t_x or t_y), the default is $t_{start} = t_s$, and hence the $\min\{t_s, t_x, t_y\}$ term in (6.16). T_{pk} and T_{win} are tuning parameters. A significant peak or trough is the extreme value compared to the surrounding signal, and the surrounding signal is defined as T_{pk} either side of the peak or trough. Extreme is defined as greater than X_{pk} for a peak or less than X_{tr} for a trough:

$$\begin{aligned} X_{pk} &= \max\{C\} \\ X_{tr} &= \min\{C\} \end{aligned} \tag{6.19}$$

where C is the set of points surrounding the potential peak or trough, defined as:

$$C = \left\{s(t_i) : s(t_i) \in R_f,\; t_s \le t_i \le t_s + T_s\right\} \tag{6.20}$$

and T_s is a tuning parameter.

The end time is described by the experts as the end of the flat region, the time when breathing movements restart (t_{end} in Figures 6.2, 6.3 and 6.4). The end time is typically clearer than the start time as there are no decay curves as sometimes occur at the start of an apnoea. Given a flat region R_f within an apnoea, the R_f end time t_e is usually not the same as the apnoea end time t_{end}. The flat region extends up to the point where the standard deviation about the centre of the window length L_f exceeds the threshold M_f, t_e in Figure 6.4. Thus, the signal within L_f must have deviated significantly so t_{end} is constrained to be within the window, i.e.: within $\frac{L_f}{2}$ of the end of the flat region ($t_e + \frac{L_f}{2}$ in Figure 6.4). The end of the apnoea is when breathing movements resume, and at this time the signal amplitude increases or decreases beyond the amplitude of $s(t)$ within R_f. The time of the start of the first significant deviation away from the flat region is the end of the apnoea. The end time may be defined as:

$$t_{end} = \min\left\{X, Y, t_e + \tfrac{L_f}{2}\right\} \tag{6.21}$$

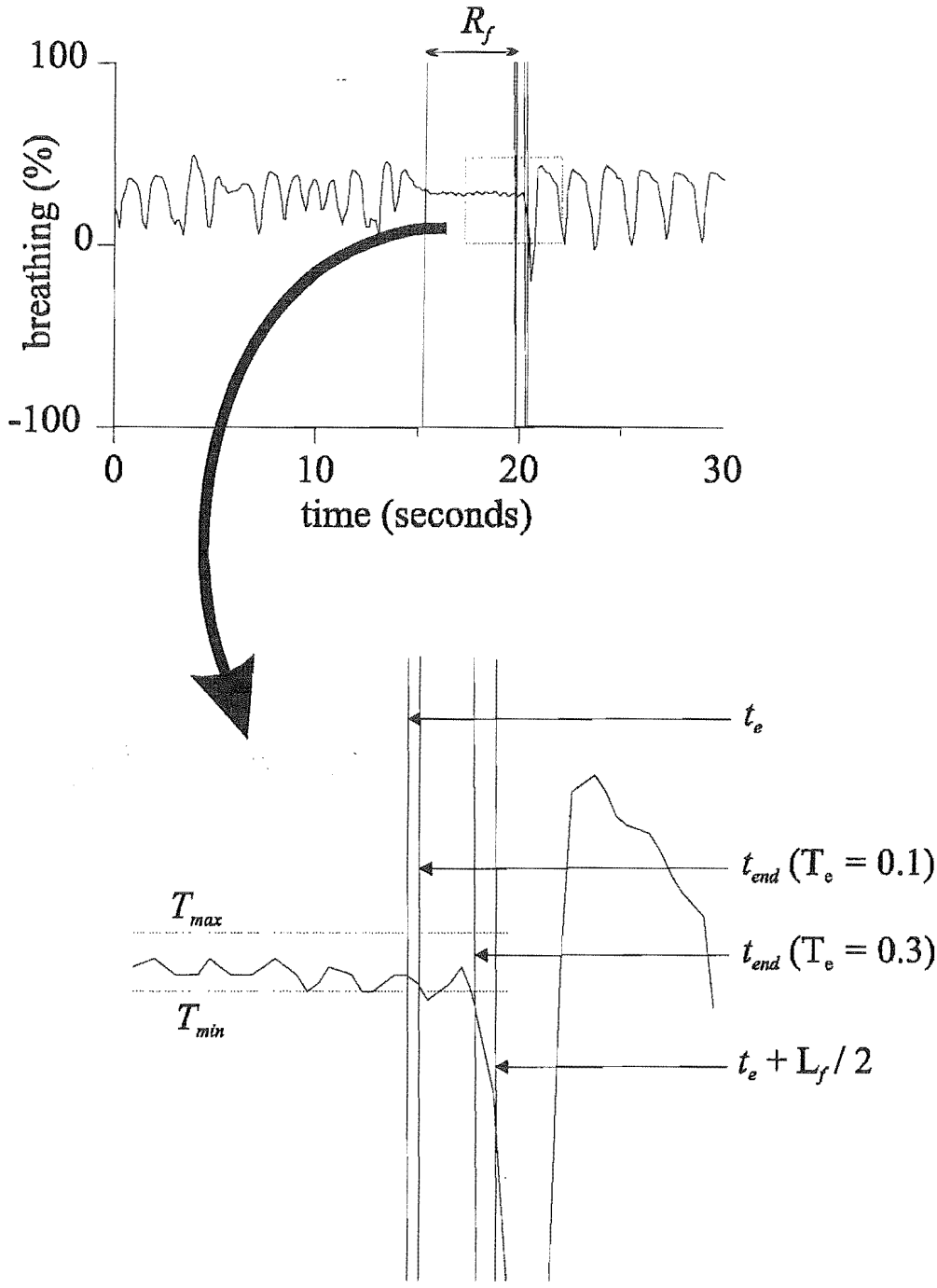

Figure 6.4 End point of apnoea with flat region R_f for $T_e = 0.1$ seconds and $T_e = 0.3$ seconds; the end of the apnoea t_{end} is a point within half a window L_f of t_e, and followed by T_e seconds of signal below $k_{min}T_{min}$ or above $k_{max}T_{max}$ (in this case $k_{max} = 0.0$ and $k_{min} = 0.0$).

where the X is the set of points of greater amplitude than flat region, defined as:

$$X = \left\{ t_x : s(t_i) > F_{max} \forall t_i : t_x \le t_i \le t_x + T_e, \quad t_e \le t_x \le t_e + \frac{L_f}{2} \right\} \tag{6.22}$$

where the times t_x are when the signal first exceeds a threshold F_{max} for a period of time T_e. Y is the set of points of lower amplitude than the flat region, defined as:

$$Y = \left\{ t_y : s(t_i) < F_{min} \quad \forall t_i : t_y \le t_i \le t_y + T_e, \quad t_e \le t_y \le t_e + \frac{L_f}{2} \right\} \tag{6.23}$$

where the times t_y are when the signal first drops below a threshold F_{min} for a period of time T_e. Figure 6.4 illustrates how a higher T_e means the apnoea end point t_{end} tends to be further towards the end of the window.

F_{min} and F_{max} are thresholds that define when the signal amplitude is significantly greater or less than the signal amplitude during the flat region, as illustrated in Figure 6.4. However, the flat regions are not always stationary, and so the mean amplitude may be increasing or decreasing from the start to end of R_f. Thus, the thresholds are relative to the end of the flat region only:

$$\begin{aligned} F_{max} &= k_{max} \max\left\{R_f \cap Z\right\} \\ F_{min} &= k_{min} \min\left\{R_f \cap Z\right\} \end{aligned} \tag{6.24}$$

where k_{max} and k_{min} are tuning parameters, and Z is the set of points at the end of the flat region, defined as:

$$Z = \left\{s(t_i) : t_i \le t_e - \tfrac{L_f}{2}\right\} \tag{6.25}$$

Using the above equations, the two measures of duration are defined as:

$$p_{21} = t_e - t_s \tag{6.26}$$

and

$$p_{22} = t_{end} - t_{start}. \tag{6.27}$$

Finally, some events are rejected outright if they are below a threshold T_{min}. Although some apnoeas could have values of p_{21} and p_{22} that are less than the minimum duration, the values are likely to be close to the minimum duration. The purpose of T_{min} is to reject all events that are definitively too short to be apnoeas.

6.2.3.2.3 Thinness

The third property of the flat region is *thinness*, the degree to which the flat region is thin compared to regions of surrounding breathing. Low amplitude signals of breathing often have flat regions with the property of flatness (high p_1), but to an expert eye these are recognised as breathing. An example is shown in Figure 6.5. Thinness is a property that distinguishes flat regions of low amplitude breathing from flat regions of apnoea. Conceptually thinness is similar to flatness, but the description of flatness does not always distinguish low amplitude signals. One type of event that has high p_1 is where the signal suddenly changes from high to low amplitude, probably due to the infant changing position (Figure 2.13) and the start of the low amplitude breathing is flat relative to the prior breathing. Low amplitude signals have different characteristics; they are not scaled down versions of normal or high amplitude signals. As seen in Figure 6.5, much of the signal is flat with each breath oscillation represented by shorter deviations and flat peaks. Miles [1989] noted that: *The most serious opportunity for mis-scoring occurs when the overall amplitude of the respiration signal is artificially small.* Apnoea monitors having false alarms due to low amplitude breathing has been a problem for many years: *...in the majority of instances* [of apnoea] *respiration had not stopped, but had become shallow* [Valdes-Dapena 1980]. As low amplitude signals are perhaps the greatest cause of false detections, thinness is developed as a separate property that specifically distinguishes between apnoeas and low amplitude breathing signals.

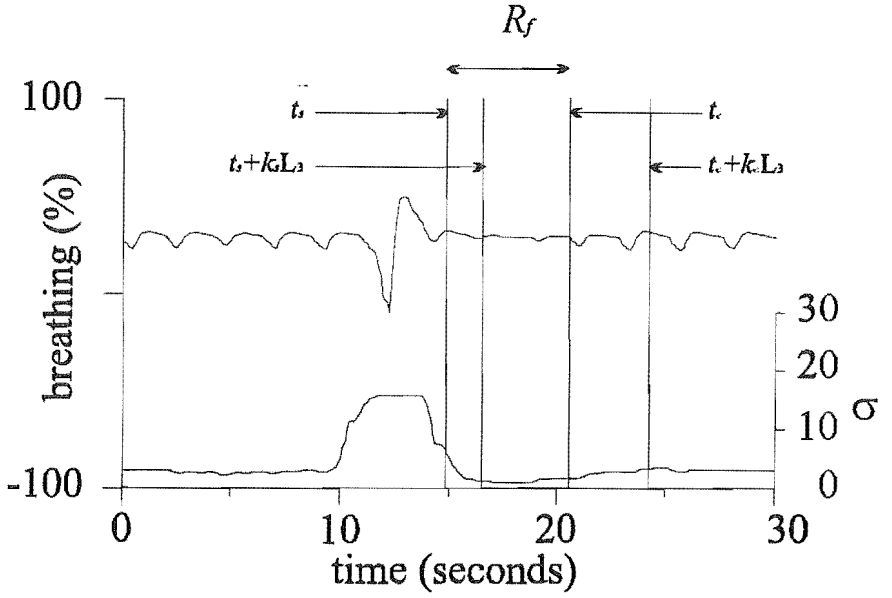

Figure 6.5 Low amplitude breathing with a flat region R_f. The standard deviation calculated with $L_3 = 3.7$ seconds is shown beneath the breathing signal. The start and end points of the window used to calculate p_3 are shown for $k_s = 0.5$ and $k_e = 1.0$.

As the standard deviation is a measure of flatness, the *change* in standard deviation is a possible measure of the thinness of the signal. Given a region of low standard deviation R_f, the greater the change in the standard deviation, the thinner the signal at that point. The greater the standard deviation of the standard deviation, the thinner the flat region relative to surrounding breathing, and the greater the thinness of the flat region. A measure of thinness p_3 is defined:

$$p_3 = \sigma\ [s_{d3}(t)]_W \tag{6.28}$$

where the window W is defined as:

$$N = (t_e + k_e L_3) - (t_s + k_s L_3) \tag{6.29}$$

and k_e and k_s are tuning parameters, and s_{d3} is the standard deviation signal:

$$s_{d3}(t_i) = \sigma[s(t_i)]_{L_3}, t_e + k_e L_3 \le t_i \le t_s + k_s L_3 \tag{6.30}$$

The standard deviation signal s_{d3} is calculated over a window L_3 at each time t_i in a region from $t_e + k_e L_3$ to $t_e + k_s L_3$.

6.2.3.2.4 Smoothness

The fourth property of the flat region is *smoothness*. The experts noted that most non-apnoea events with high measures of p_1, p_2 and p_3 are characterised by uneven, irregular flat regions. For example, periods of flat signal may be followed by a few oscillations and interspersed with one or two larger deviations, as in Figures 6.6 and 6.7. Irregular signals have been noted to cause problems in detecting apnoeas previously [Jeffrey, et al. 1981]. As the opposite of these characteristics, the property of smoothness is ascribed to apnoeas. Smoothness refers to how smooth the flat region is, or how regular or even any variations are. A smooth flat region is either flat or contains regular oscillations, similar to those corresponding to cardiac movement. There are two causes of non-apnoea events with low smoothness:

1. The infant has paused breathing and taken a very small breath, in or out, then paused again, and then restarted breathing (as in Figure 6.6 and also Figure 3.1); and

2. A poor signal with low amplitude and a time of irregular, disturbed breathing, which could be due to perturbed breathing or a noisy signal (as in Figure 6.7).

Both types of events cause false detections.

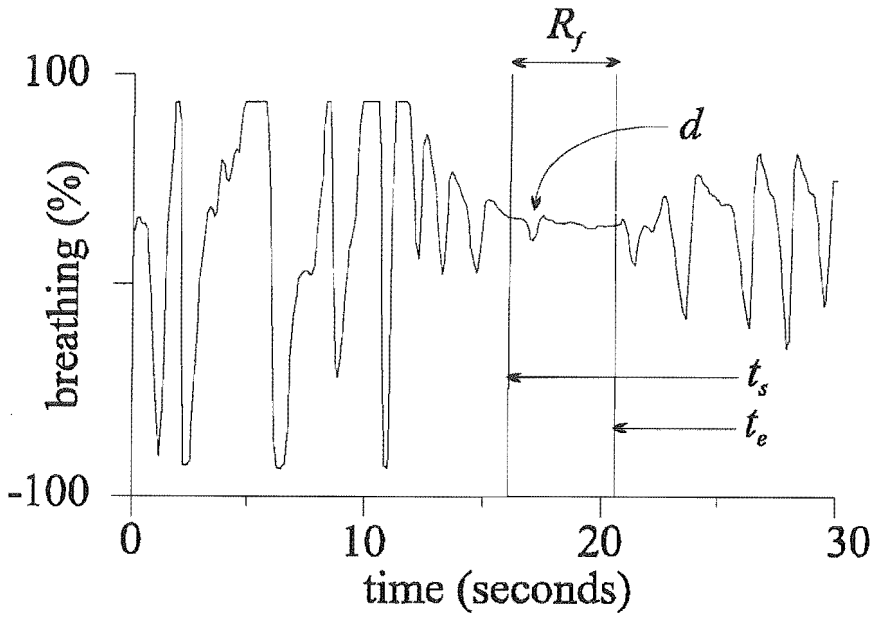

Figure 6.6 Non-apnoea with deviation d in flat region R_f corresponding to small breath.

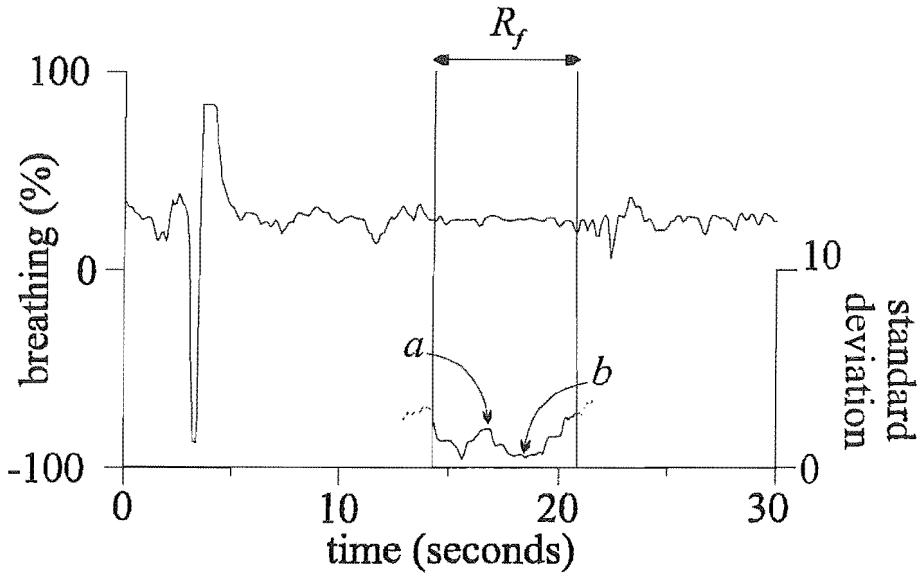

Figure 6.7 Non-apnoea event showing uneven flat region and varying standard deviation, with maximum at a and minimum at b.

There are two signal characteristics that distinguish smoothness. Firstly, considering the first type of non-apnoea event mentioned above, the flat region is interrupted with a single oscillation, and then returns to being flat. Secondly, considering the second type of non-apnoea event mentioned above, there are very flat parts of the region interspersed with rapid amplitude changes or sharp peaks. Therefore, two separate measures are used to defined smoothness.

Considering the first type of event, there is one part of the flat region that is significantly different to the remainder of the flat region. Hence, comparing the standard deviation of the oscillation with the standard deviation of the remainder of the flat region gives a measure of

smoothness. The standard deviation of the remainder of the flat region is given by the median of the standard deviation as calculated at each point on the flat region, and the standard deviation of the oscillation could be given by the maximum standard deviation of any point within the flat region. A measure of smoothness p_{41} is defined:

$$p_{41} = \frac{y}{z} \tag{6.31}$$

where z is the standard deviation of the oscillation, defined as:

$$\begin{aligned} z &= \max\{S\} \\ S &= \left\{ s_{d4}\left(s(t+1)\right) - s_{d4}\left(s(t)\right) : s \in R_4 \right\} \end{aligned} \tag{6.32}$$

and y is a measure of the standard deviation of the remainder of the flat region, defined as:

$$\begin{aligned} y &= \text{med}\{R\} \\ R &= \left\{ s_{d4}\left(s(t_x)\right) : s(t_x) \in R_4 \right\} \end{aligned} \tag{6.33}$$

and where s_{d4} is the standard deviation signal calculated over a window L_4:

$$s_{d4}(s) = \sigma[s(t)]_{L_4} \tag{6.34}$$

and p_{41} is defined over a region R_4:

$$R_4 = \left\{ s(t_x) : t_{start} + \mathrm{T}_{\text{shiftst}} \le t_x < t_{end} + \mathrm{T}_{\text{shiftend}} \right\} \tag{6.35}$$

where $\mathrm{T}_{\text{shiftst}}$ and $\mathrm{T}_{\text{shiftend}}$ are tuning parameters that control exactly where the region R_4 is located.

Considering the second type of event, there may be several rapid changes in amplitude or oscillations as opposed to one isolated deviation. The distinguishing characteristic is the difference between the very flat parts of the region and the sharp oscillations that are large relative to cardiac oscillations. Hence, the difference between the maximum and minimum standard deviation of the flat region points gives a measure; two such points are a and b are shown in Figure 6.7. A second measure of smoothness p_{42} is defined:

$$p_{42} = \frac{\min\{Q\}}{\max\{Q\}} \tag{6.36}$$

where Q is the set of standard deviation values at points in, R_4 defined as:

$$Q = \left\{ s_{d4}\left(s(t_x)\right) : s(t_x) \in R_4 \right\} \tag{6.37}$$

where the parameters are the same as for p_{41}.

6.2.4 Optimisation Results

Having defined the properties, they are further refined by optimising the tuning parameters. As defined, the properties could be applied to a variety of breathing signals, but in order to test them, they need to be evaluated relative to a reference set of breathing signals and apnoeas.

All property parameters are tuned to optimise SI(5) as in (6.1). The value of L_1 in (6.14) is adjusted to maximise SI(5) for B_1, giving $L_1 = 0.9$ seconds (see Table 6.1). The smaller L_1, the more sensitive the measure of standard deviation is to small variations in the signal. In the context of five second apnoeas and 1.9 second flat regions, a 0.9 second window means the standard deviation is relatively sensitive.

The optimum values of the tuning parameters are shown in Table 6.1. A grid optimisation was performed, meaning that every combination of parameter values over appropriate ranges was optimised. The values for each parameter range from the minimum to the maximum incremented by the step, as shown in the three right side columns of Table 6.1. The flat region tuning parameters were tuned to optimise the performance as measured by P as in equation (4.1), the duration measurement was tuned to minimise the MSE between the calculated and actual start and end points, and the remainder of the properties were tuned to optimise the discrimination as measured by SI(5) as in equation (6.1).

The standard deviation window L_3 is 3.5 seconds, with the longer window length compared to other windows lengths used for calculating standard deviations reflecting the fact that thinness is a property that is based on a wider context than flatness, where $L_1 = 0.9$ seconds, or smoothness where $L_4 = 0.3$. The tuning parameters shift the window over which p_3 is calculated, with $k_s = 1.05$ and $k_e = 0.65$. $L_4 = 0.3$ seconds, and so the standard deviation is sensitive to small deviations, reflecting the fact that smoothness is a property related to small changes on the flat region.

For most parameters, the performance is not greatly affected if the parameter value is slightly shifted from the optimum. The effects of parameter values on performance are illustrated by plotting the performance figures for each parameter value. For example, a display of the results of tuning the flat region window L_f is shown in Figure 6.8. Note that the values of 0.7, 0.9, 1.1, 1.3 and 1.5 all have similar minimum and median penalty values, whereas higher values of L_f result in greatly increased median penalty values. Thus, based on the median, values of L_f below 1.5 are more likely to result in optimum performance. Different results can be seen for the duration threshold T_r, displayed in Figure 6.9. The value of T_r which has the minimum penalty

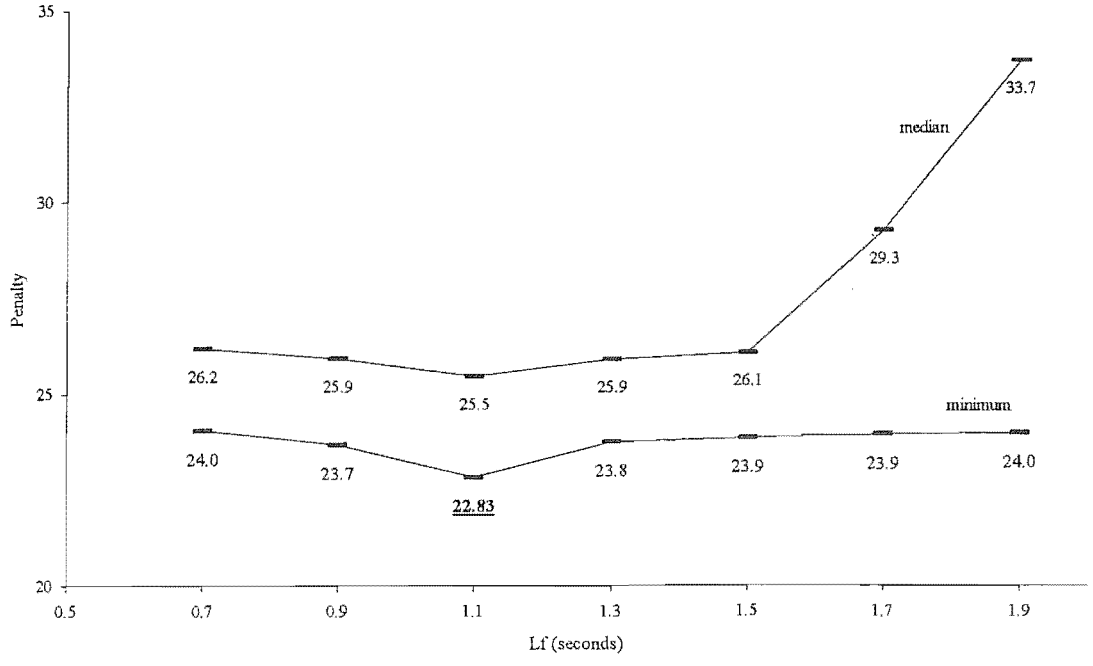

Figure 6.8 Penalty values for various window lengths. For each setting of L_f, there are approximately 2500 combinations of other parameters, resulting in approximately 2500 Penalty values. The minimum and median values are shown for each setting of L_f.

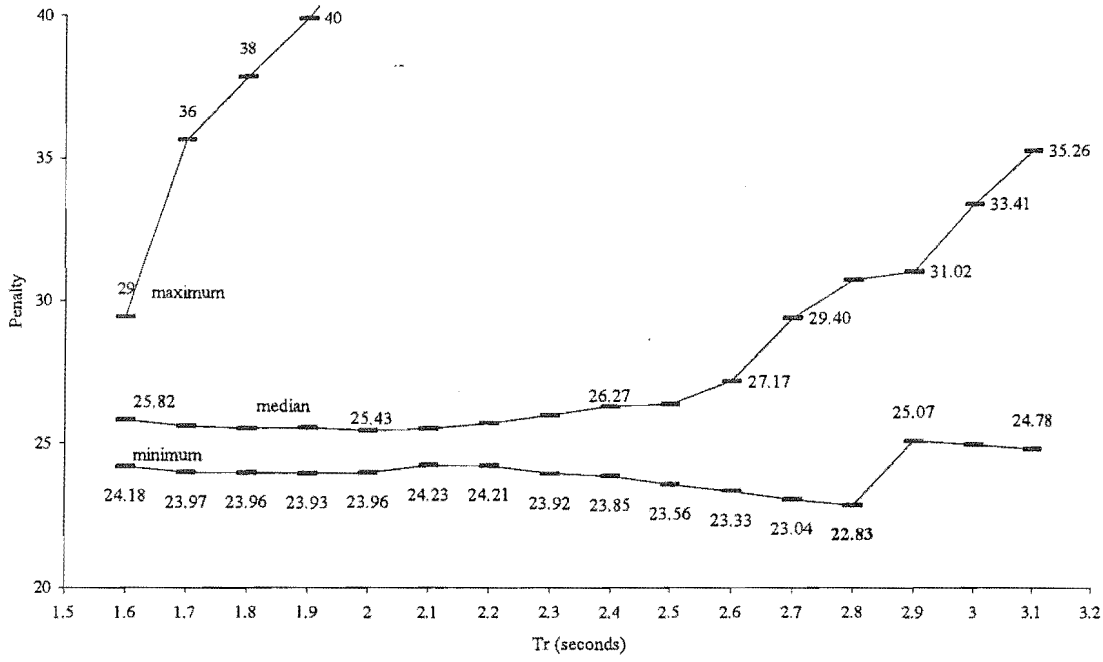

Figure 6.9 Performance for range of values of flat region duration threshold T_r.

does not have the minimum median penalty, but for values of T_r up to 2.2 seconds, there is little significant difference in performances.

A further example of the optimisation of a parameter is illustrated for the tuning parameter k_f. In this case, there is a more defined optimum setting with a clear minimum, as illustrated by Figure 6.10. The remaining parameters display similar effects on performance as those already illustrated in Figures 6.8, 6.9 and 6.10.

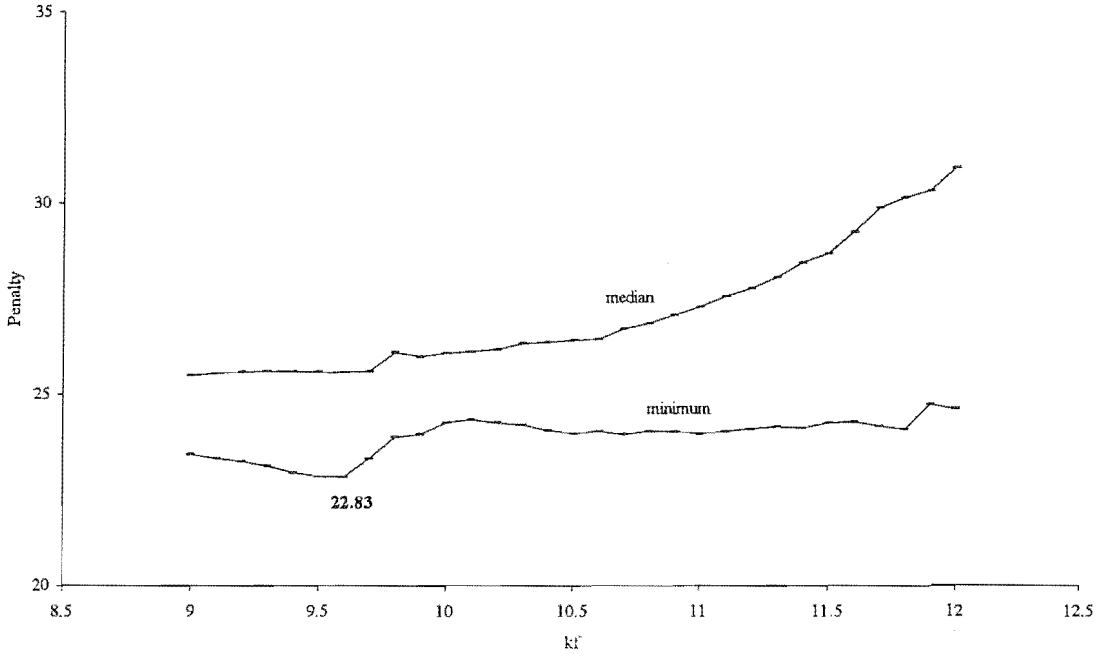

Figure 6.10 Penalty values for settings of the tuning parameter k_f.

6.2.5 Discriminating Power and Independence of the Properties

The properties presented in the previous section were evaluated. They were tested for discrimination using a different measure of discrimination than the specificity index, and they were also tested for independence.

Figure 6.11 shows different cumulative frequency curves for apnoea and non-apnoea events. As explained in Section 6.2.2.4, if the two cumulative frequency curves of a property measure for apnoea and non-apnoea events are different, then that measure has discriminating power (see Section 6.2.2.3). The Smirnov test is based on the maximum separation D between distributions (defined by (6.2)), and in all cases the p value is very close to zero, confirming the discrimination power [Lindgren 1976]. Note also that the distributions of test data (B_2 as described in Section 5.2.3) shown in Figure 6.11 are close to the distributions for the training data for most properties. For property p_3, B_1 and B_2 non-apnoea distributions are the same, but the apnoea distributions differ significantly, with the B_2 values being lower on average. This is caused by the low amplitude nature of the B_2 signals [Tappin, et al. 1997]: many apnoeas occurred during lower amplitude signals, resulting in a lower thinness. Overall, the B_2 distributions are very close to the B_1 distributions, confirming that the properties are indeed *general* properties.

Note that the non-apnoea events are false positives from the system presented in Section 5.1, and these non-apnoea events are similar to apnoeas. In other words, the previous system does not have further discriminating power. Therefore, the fact that the properties measures outlined in this chapter discriminate between any apnoea and non-apnoea events means that they are more effective than the measures used by the previous system. Hence, these property measures successfully discriminate between apnoea and *similar* non-apnoea events.

The power of discrimination varies for the properties. For properties p_1, p_{21} and p_3, the power of discrimination is relatively high, reflecting the fact that those properties have good general discriminating power. For the other properties the power of discrimination is relatively low. Properties p_{41} and p_{41} are designed to discriminate between a smaller number of events and consequently they have less overall discrimination. For the duration p_{22}, the property is optimised to maximise accuracy, not discrimination. The overlap in the curves means that p_{22} does not give a direct measure of apnoea likelihood, but there is a definite difference between the curves and therefore there is some discrimination.

Table 6.2 shows the correlation coefficients for the combinations of the properties. The six measures of the four properties all represent different information, with some properties being relatively independent, whereas others are partially correlated.

Property	p_1	p_{21}	p_{22}	p_3	p_{41}
p_{42}	0.15	0.32	0.04	0.11	0.22
p_{41}	0.48	0.34	0.12	0.40	
p_3	0.40	0.35	0.07		
p_{22}	0.15	0.50			
p_{21}	0.57				
Average				0.28 ± 0.16	

Table 6.2 Correlation coefficients between all properties.

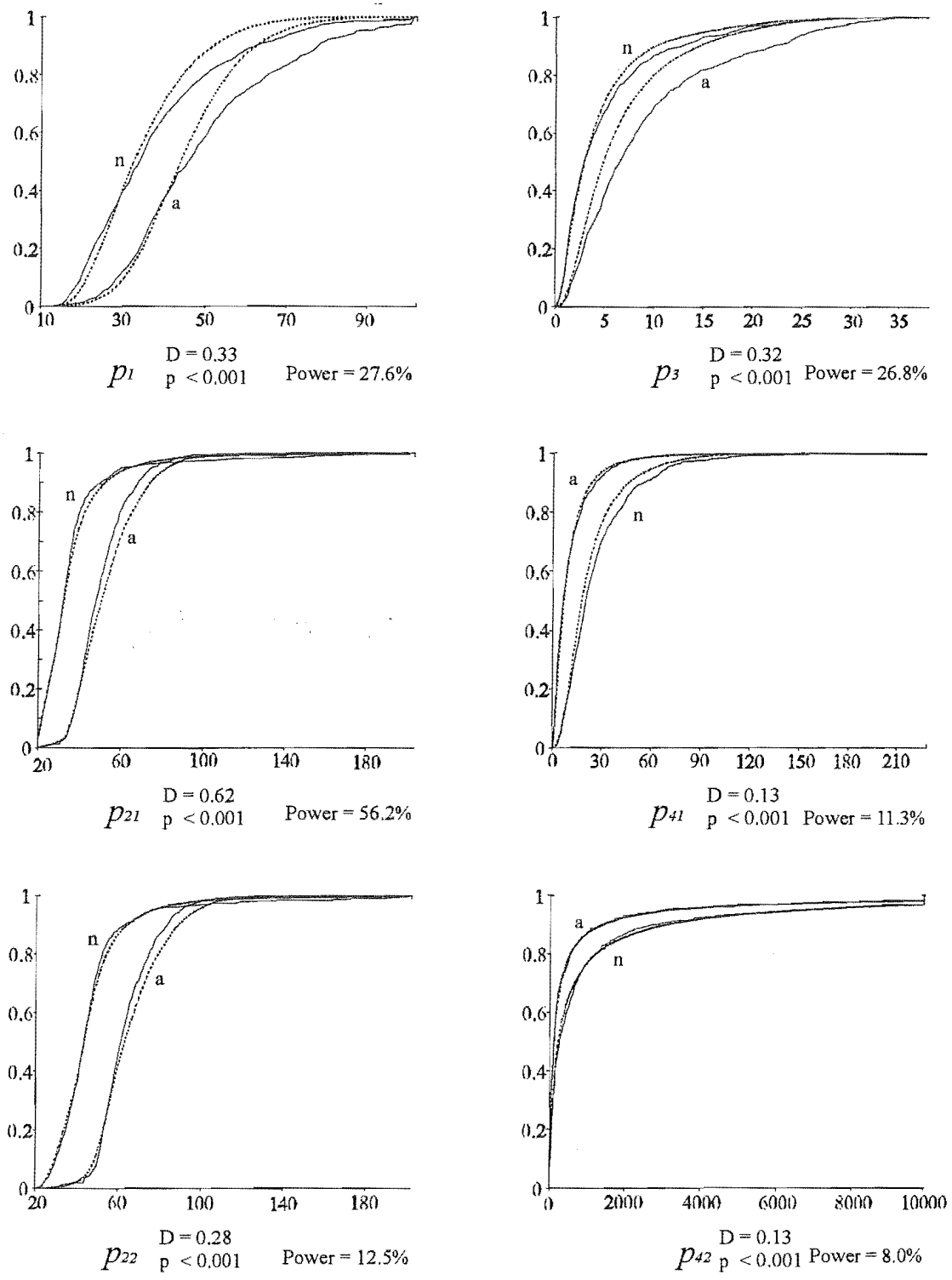

Figure 6.11 Normalised cumulative frequencies of all apnoea (a) and all non-apnoea (n) events for each of the six property measures. The x axis represents the value of the measure, and the y axis represents the proportion of events. Any point on a graph is the proportion of events that have the property measure of that value or less. The solid lines are for the training events from B_1, and the dashed lines are for the test events from B_2. The measure D and the power are displayed, and are calculated as described in the text.

The highest correlation of 0.565 is between flatness p_1 and flat region duration p_{21}, as the flatter a signal the longer it is likely to be below a threshold. The next highest correlation of 0.50 is between flat region duration p_{21} and apnoea duration p_{22}, and this is expected as the apnoea duration is calculated from the flat region duration. Flatness p_1 and smoothness p_{41} are also related, with the flatter a signal the smoother it is. However, correlation coefficients of 0.50 and 0.57 still indicates that two properties, whilst not independent, do measure different information.

The lowest correlation is between p_{22} and p_{42}, duration and the smoothness. These two measures are therefore almost independent. Whilst it was possible that some of the measures that had similar descriptions would be highly correlated, this turned out to not be the case. The correlation of 0.15 between the two measures of duration confirm that each measures different information, and similarly with the two measures of smoothness.

6.3 Classification of Breathing Signals

This section describes an expert system for apnoea detection. The system detects possible apnoeas and then measures their properties as described in Section 6.2.3, and analyses the measures in order to classify the events. The system is trained on a variety of signals.

6.3.1 Objectives

There are two reasons for developing an expert system. The main reason is to have a reliable detection system that emulates a human expert. A second reason, in the context of this research, is to test the validity of the properties described in Section 6.2.

Even though the system is designed and tested with a limited number of signals, the aim is that the system may be successfully applied to other types of signals. Breathing signals recorded from different instruments are often similar, as seen in Figure 2.4, and therefore it is likely that a system that works with one type of signal would need only minor adjustments to work with most other signals. As well as signal type, expert opinion is a variable: it is possible that the reference expert opinion of what constitutes an apnoea could change, either because a new expert is using the system or because their opinion has been revised. If the system can be retrained, and the properties tuned again, then new signals and new reference apnoeas can be adapted to. Therefore, an aim is to develop the system to be trainable, describing the training in a systematic manner that can be reproduced.

A system is considered reliable if it detects apnoeas in a variety of signals, relative to a variety of experts. In other words, the system could be used on a new recording with the confidence that it accurately detects apnoeas as determined by an expert other than the one used for the training data, with known approximate false negative and false positive rates. Another indication of reliability is the repeatability of results, and the consistency of results across different recordings. The advantage of an algorithm compared to human experts is that results are repeatable and consistent.

As mentioned in Section 1.4, one of the objects of this research is to objectively describe the methodology of detection. Having a single "black box" algorithm is not as useful as having a sequence of well defined, logical steps. Calculating the deterministic properties is essentially one stage, but more are needed for a complete system. There is a potential trade-off between clarity

and accuracy, as the sequence of steps may constrain the system to be less than optimum. Summarising, the objectives are that the system should:

1. achieve improved apnoea detection performance compared to previous systems;
2. be reliable for different signals and relative to different experts' opinions;
3. be trainable;
4. be composed of clearly defined, logical steps; and
5. measure properties p_1 to p_4 as described in Section 6.2.3.

6.3.2 System Design

The basic requirement of any apnoea detection system is that given a breathing signal as input, a list of apnoeas and their durations is produced. Usually, certain properties of the raw signal are measured as a means of extracting vital information whilst also reducing the size of the information. The properties are then analysed to classify breathing. This section describes an analysis and classification system.

6.3.2.1 Overview

A system S as in Figure 6.12 (a) is defined, where S receives as input a breathing signal $s(t)$ and produces a list of apnoeas $\underline{A}$. The system is designed to be flexible by being able to be retrained with different training data generated from other signals, other property measures, other experts, or an expert with a revised opinion.

The system is initially split into two stages, as in Figure 6.12 (b). The first stage S_1 produces a list of possible events with measures of properties that characterise the presence of an apnoea. The property measures of the events are passed to the second stage S_2 that classifies events as apnoea or non-apnoea. This division approximately models expert detection in that human experts quickly scan through a breathing signal until they notice a region with apnoea-like properties; they then evaluate that region in detail and classify it as apnoea or non-apnoea (see Section 3.3). Extracting salient information is thereby separated from classification, and classification is based on a small number of property measures as opposed to all the raw data.

6.3.2.2 Property Measurement

The first stage S_1 produces measures of properties of possible apnoeas. It is further split into two stages. Firstly, events that are possible apnoeas are detected, S_{1f} in Figure 6.12 (c), and secondly, properties of the signal at those events are measured, S_{1p} in Figure 6.12 (c).

Most breathing signal is not apnoea—it is clearly distinguishable as breathing. Hence the first stage S_{1f} aims to eliminate signals that clearly correspond to breathing. By analysing the continuous breathing signal $s(t)$ and producing a list of possible apnoea events, the output of S_{1f} is a sequence of regions of $s(t)$. The detection task is then reduced to evaluating a number of events as opposed to a continuous recording.

As detailed Section 6.2.3, an apnoea is essentially a flat region, and if only events consisting of flat regions are detected then most breathing is excluded. As S_{1f} is a collection stage, it is important that very few apnoeas are excluded. Therefore, S_{1f} detects flat regions in the signal using the algorithm described in Section 5.1.1, which detects almost all apnoeas but which also has a high rate of false positives. However, if *all* apnoeas are detected then there may be large

numbers of non-apnoea events detected, and therefore the S_{1f} stage can exclude a small number of apnoeas, where small is set at less than 1% of all apnoeas.

The deterministic properties of the events detected by S_{1f} are measured by S_{1p}, shown in Figure 6.12 (c), and described in Section 6.2.3. The measures p_1 to p_4 are passed to the classification sub-system.

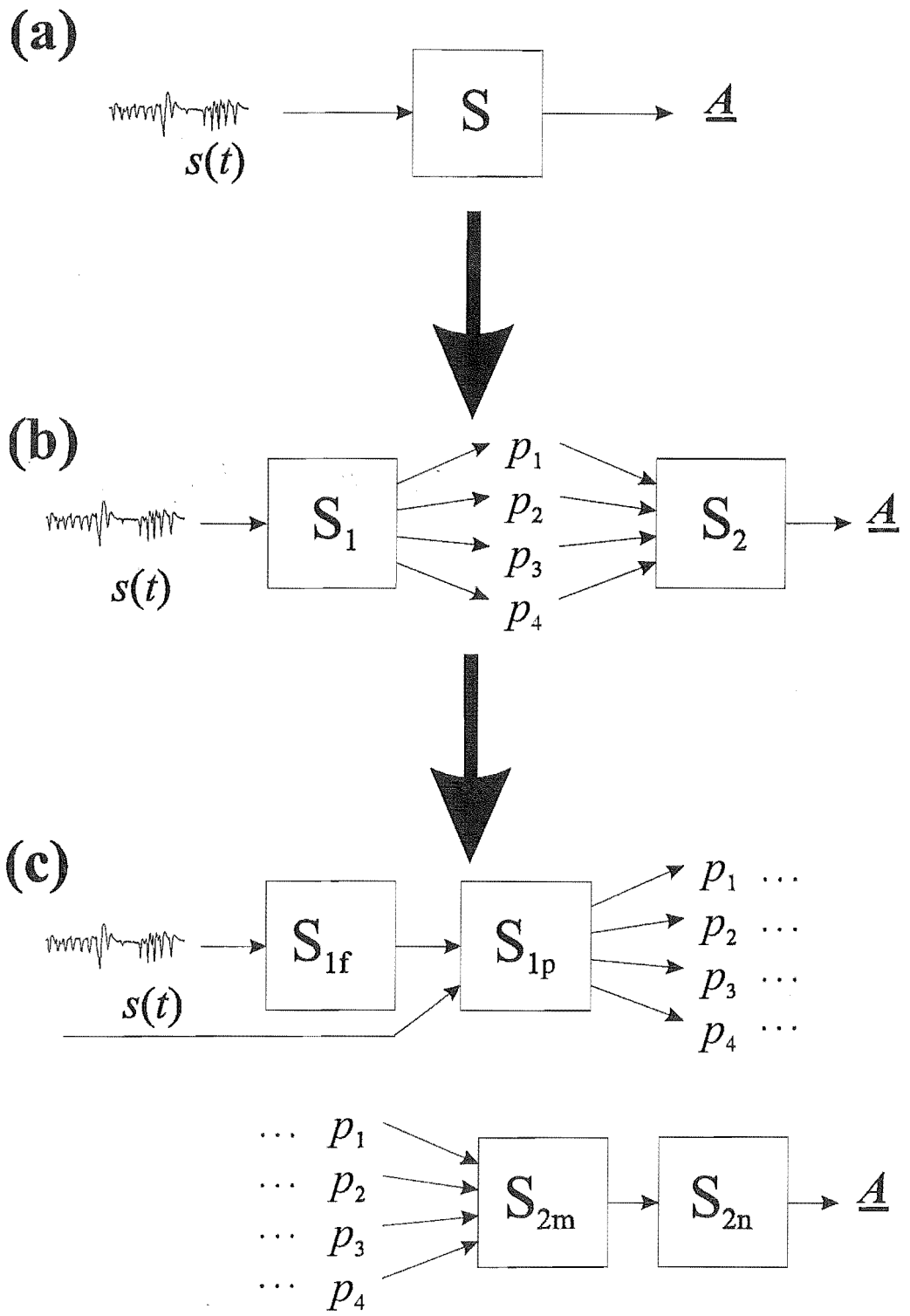

Figure 6.12 (a) General detection system S that calculates a list of apnoea $\underline{A}$ from input breathing signal $s(t)$; (b) overall system design with S_1 measuring properties p_1, p_2,...etc. of possible apnoea; properties passed to S_2, which classifies events as apnoea or non-apnoea; and (c) S_1 and S_2 further divided as described in text.

6.3.2.3 Classification

S_2 is a classification stage. Given an event that is a possible apnoea, S_2 uses the measures of properties from S_1 to evaluate the likelihood of the event being an apnoea. A neural network approach is used, as neural networks have been shown to be effective in pattern recognition and classification problems [Lippmann 1987]. A raw breathing signal has been analysed with a large neural network (around 150 nodes), but the only successful trainings were on small data sets, and the network did not generalise [Sturman 1991]. With the present system, large amounts of data are condensed, and the classification task of S_2 is simple compared to analysing the entire breathing signal. S_2 is split into two stages: translating property measures into a format for input to the neural network, and the neural network itself.

Inputs to Neural Network

The neural network stage (S_{2n} in Figure 6.12 (c)) is the actual evaluation stage. The criterion for inputs to the neural network is that they are normalised, in this case between 0.0 and 1.0, and therefore the first stage is mapping property measures to appropriate input values (S_{2m} in Figure 6.12 (c)) [Pao 1989]. The measures of a property are mapped to input values between 0.0 and 1.0 using a mapping function g_x:

$$i_x = g_x(p_x) \tag{6.38}$$

where p_x is the measure of property x and i_x the input for property x. The convention used is that as i_x approaches 0.0 the less apnoea-like is the property, and as i_x approaches 1.0 the more apnoea-like is the property. The function g_x is constrained to be monotonically increasing or decreasing, and is therefore a one-to-one mapping that preserves the characteristics of the distribution of p_x. For events that are clearly apnoea or non-apnoea, the exact value of i_x is not important if it is set close to 1.0 or 0.0 respectively. By maximising the separation between the i_x values of similar apnoea and non-apnoea events, referred to as *boundary* events, the differences between boundary apnoea and non-apnoea i_x values are emphasised. Therefore, while also normalising the inputs, the mapping g_x spreads the i_x values of boundary events across the [0,1] range.

The general form of g_x must map a range of [-∞,∞] to [0,1] so that all possible values of p_x can be mapped to within the desired range. A type of function that has those characteristics is a *limiter* function; for example the function g:

$$g(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6.39}$$

However, the mapping ideally maintains the separation between boundary events, which requires different characteristics to those of a limiter function. The inverse tangent is one function that has all the desired characteristics: it is monotonically increasing, it maps [-∞,∞] to a finite range, and it has an approximately linear region about a domain of zero. Therefore, the inverse tangent is adapted to a suitable mapping function g_x.

Initially consider a function $g(p_x)$ where:

$$g(p_x) = \mathrm{Tan}^{-1}(p_x) \tag{6.40}$$

g maps the range [-∞,∞] to $[\frac{-\pi}{2}, \frac{\pi}{2}]$, and so to map to the range [0,1] g can be modified to the following:

$$g(p_x) = \frac{\mathrm{Tan}^{-1}(p_x)}{\pi} + 0.5 \tag{6.41}$$

The range of p_x is likely to be finite, and the range of p_x for the boundary events is likely to be significantly less than the entire range of p_x. In order to maximise the separation between boundary events, the approximately linear region of g can be narrowed or widened to cover the entire p_x range of boundary events. The resulting g needs to be rescaled to ensure the entire [0,1] range is used. Two additional parameters are therefore included:

$$g(p_x) = \frac{\mathrm{Tan}^{-1}(k\,p_x)r}{\pi} + 0.5 \tag{6.42}$$

where k controls the width of the approximately linear region, and r controls the extent of the [0,1] range that is used. If the boundary events do not occur in the centre of the range of all events, g can be shifted and hence more parameters are required:

$$g(p_x) = \frac{\mathrm{Tan}^{-1}\left(k(p_x - sx)\right)r}{\pi} + sy \tag{6.43}$$

where sx and sy shift the x and y axes of g respectively. If the range of p is finite and $sx \neq 0$, then g is shifted on the y axis so the parameter sy shifts g back. Equation (6.43) allows the mapping to be adjusted to suit a variety of ranges of p_x and a variety of ranges of boundary event p_x.

A final refinement is added. The parameters k, r, sx and sy in (6.43) vary for different p_x. If p_x is normalised, then the parameters can be compared between mapping functions for different properties. Therefore, the form of g_x used is similar to g in (6.43) but with some modifications to the parameters:

$$g_x(p_x) = \frac{\mathrm{Tan}^{-1}\left(2k_x\left(\frac{(p_x - m_x)}{2A_x} - sx_x\right)\right)r_x}{\pi} + sy_x \tag{6.44}$$

where the parameters k_x, r_x, sx_x and sy_x control the exact form of the mapping, and their default values are 10, 1.0, 0.5 and 0.5 respectively. The terms m_x and $2A_x$ refer to the mean and the range of the measure p_x; in other words, the measure is initially normalised to between 0.0 and 1.0. The form of g_x in (6.44) can be adjusted so that the boundary event i_x values are separated as much as possible, depending on whether the boundary event p_x values occur across the whole range of p_x, a narrow central range, or an offset range.

The mapping in (6.44) as opposed to (6.43) allows the values of the parameters k_x, r_x, sx_x and sy_x to be interpreted and compared between g_x for different properties. Figure 6.13 illustrates how the tuning parameters of g_x alter the mapping. The parameter k_x controls the width of the approximately linear region so that if the boundary events occur across a small range, k_x is large and the function characteristics tend towards a limiter type function (graph a in Figure 6.13). The parameter r_x is a gain so that given a k_x, the output of g_x can be spread across the entire 0.0 to 1.0 range (graph b in Figure 6.13). Thus for low k_x, r_x increases. If the boundary event p_x values do not occur about m_x, the mapping can be shifted along the p_x axis using the parameter sx_x (graph c in Figure 6.13). If $sx_x \neq 0$, then g_x is also shifted on the i_x axis and the parameter sy_x can

compensate by shifting the mapping back along the i_x axis (graph d in Figure 6.13). The four parameters allow the mapping to be adapted to a variety of distributions of p_x to give an appropriate i_x distribution.

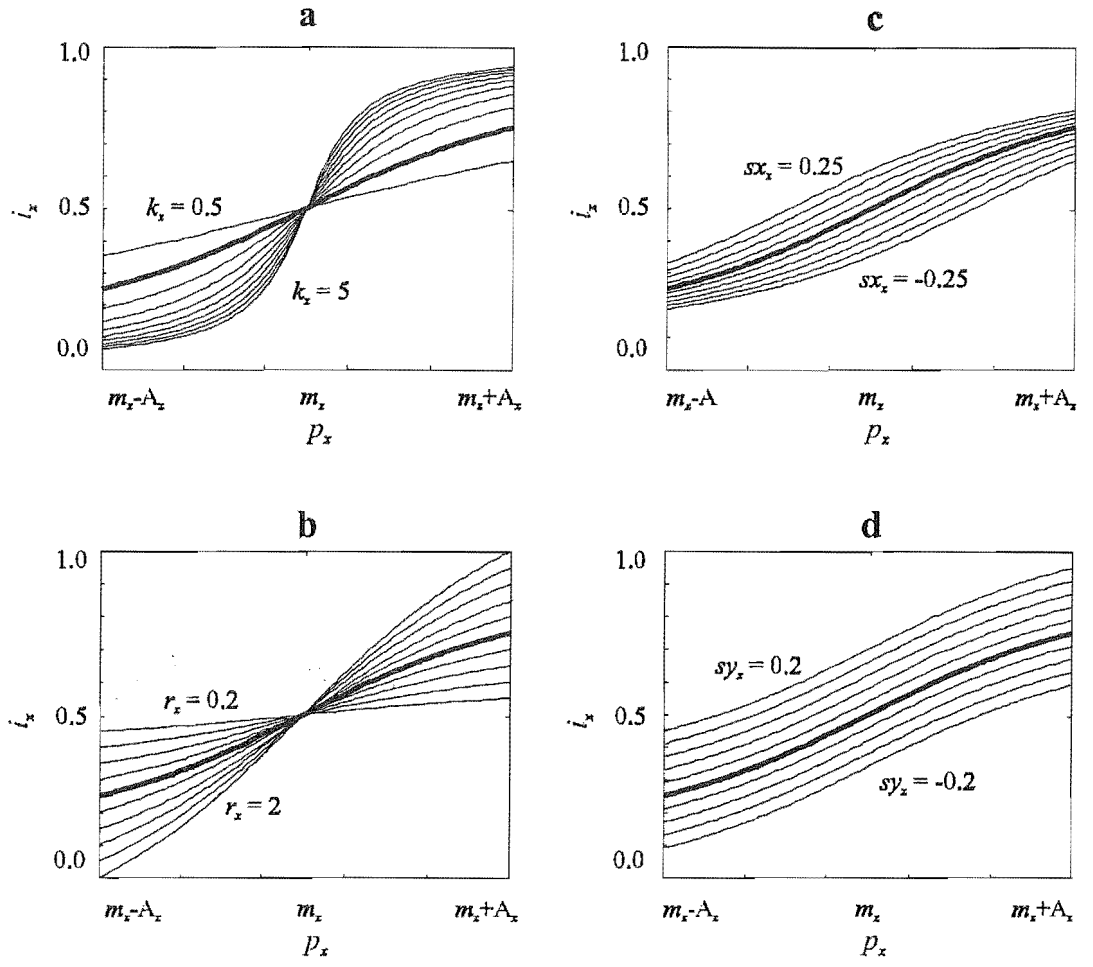

Figure 6.13 Mapping function g_x in different forms according to the tuning parameter settings. The thick centre line in all graphs is the reference g_x for k = 1.0, r = 1.0, sx = 0.0 and sy = 0.0. The graphs a, b, c and d illustrate the how the parameters k, r, sx and sy affect g_x, with various g_x plotted for incremental parameter values either side of the reference g_x.

Examples of possible resulting g_x are shown in Figure 6.14. If the boundary events are spread across the entire range of p_x, k_x is small and r_x is greater than 1.0, curve a in Figure 6.14. Otherwise, if the boundary events are located within a small part of the whole range, then k_x is large and $r_x \approx 1.0$, curve b in Figure 6.14. If the boundary events do not occur about m_x then $sx_x \neq 0.0$ and the mapping is shifted, with sy_x compensating for the i_x axis shift, as with curve c in Figure 6.14. In each case the basic form is retained.

Neural Network

The neural network classifier is a standard feedforward multi-layer network, trained with the backpropagation algorithm [Lippmann 1987, Pao 1989, Pal and Mitra 1992]. Although there have been many developments in the field of neural networks, for practical applications the standard feedforward multi-layer network architectures and backpropagation training algorithms are widely used [Chakrabarti et al. 1995, Kim and Nam 1995, Michalopoulou et al. 1995, Nekovei and Sun 1995, Moody and Antsaklis 1996].

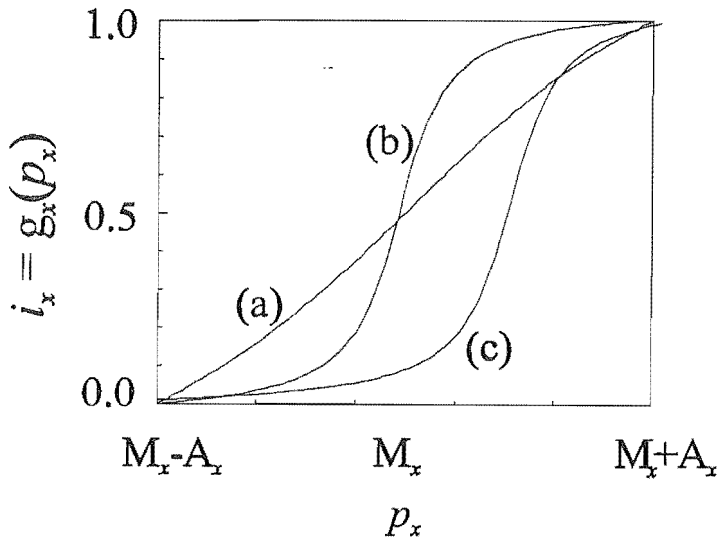

Figure 6.14 Transform function g_x which is tuned to map property p_x to input i_x over a 0.0 to 1.0 range. The boundary events are ideally spread across the approximately linear range of g_x. The different g_x are obtained using different tuning parameter values: (a) $k_x = 1.0$, $r_x = 2.0$, $sx_x = 0$, $sy_x = 0$; (b) $k_x = 7.0$, $r_x = 1.1$, $sx_x = 0$, $sy_x = 0$; (c) $k_x = 7.0$, $r_x = 1.1$, $sx_x = 0.2$, $sy_x = 0.025$.

Network and training parameters were based on previous tests (see Section 4.6). A tan-sigmoid transfer function using a bias weight and two hidden layer networks were used, along with momentum and learning rate parameters as specified in Section 4.6. The output of the neural network is a likelihood measure that is normalised to between 0.0 and 1.0. For apnoea scoring, where a list of apnoeas is required, events with a likelihood above a threshold are classified as apnoea, and all other events as non-apnoea.

6.3.3 Training Criteria

In order to train the system, training criteria are required. A full global optimisation is impractical as there are more than 30 tuning parameters in the first three stages, and the neural network is designed to be trained independently. Hence each stage is optimised to a local performance criterion.

6.3.3.1 Overall System

Since the detection of apnoea is the main concern, the system performance criterion is based on reducing missed apnoeas and reducing false detections. Given that three experts had only a 90% agreement on what signals represented apnoeas, a detection system could not agree with all human experts [Macey, et al. 1995]. The performance goal for the expert system is therefore to detect apnoeas that would be found by *any* expert, leading to an inherent high false positive rate relative to any individual expert's opinion. Although ideally all apnoeas would be detected, it is acceptable with apnoea scoring to miss a few apnoeas in order to reduce the number of false detections.

In a previous study, the agreement between two experts was measured as 97% to 98% [Kahn, et al. 1988]; in this case the experts analysed three breathing signals including airflow in conduction with other signals, such as heart rate and oxygen saturation [Kahn, et al. 1988]. In other words, even in this ideal situation, there was disagreement on over 2% of events. The

agreement of experts analysing a single breathing signal that is not airflow is therefore likely to be less than 97%. Given that under ideal circumstances 2-3% of events are still not agreed on, the system is not constrained to detect the extreme 2% of events. Hence minimising the false positives for a 2% false negative rate is used as the overall performance criterion for the system.

As mentioned earlier, the Specificity Index SI(f_n) from (6.1) is a measure of performance defined for a false negative rate f_n. Thus SI(2) is set as the overall system performance criterion, and the local performance criteria are developed.

6.3.3.2 Properties

The aim initially is to detect all possible apnoea events so as to include almost all (> 99%) apnoeas, whilst excluding as much breathing signal as possible. A performance measure to judge which of two performances is best has been described in Section 4.2.3, and is used to optimise the settings for S_{1f}.

The stage S_p is tuned by optimising the property measures, as described in Section 6.2.3. The properties provide a small number of measures that discriminate between apnoea and non-apnoea events, and therefore the optimisation criterion for S_{1p} is to maximise the Power of Discrimination, which can be measured by SI(f_n) in (6.1) for $f_n \geq 5\%$ [Mood, et al. 1974]. Hence, maximising SI(5) is the training criterion for each of the property measures p_x.

6.3.3.3 Neural Network Input Transformations

In order to set g_x to ideally map the boundary events, the parameters k_x, r_x, sx_x and sy_x are tuned to maximise the separation between the apnoea and non-apnoea distributions of i_x. Hence, a measure of separation is needed.

The separation is the difference between apnoea and non-apnoea values of i_x. The set of input values i_x for the training data can be split into two: i_{xA} and i_{xN}, which are defined as the values of i_x for apnoea and non-apnoea respectively. A measure of separation is developed by considering the ideal case where $i_{xA} = 1.0$ and $i_{xN} = 0.0$ for all events. To evaluate how close the i_x values are to the ideal, the MSE of i_{xA} with respect to 1.0 being the desired value and the MSE of i_{xN} with respect to 0.0 being the desired value can be calculated. The sum of the normalised MSE's is then a measure of separation:

$$m = \frac{1}{N_A}\sum i_{xA}^{\,2} + \frac{1}{N_N}\sum \left(1.0 - i_{xN}\right)^2 \qquad (6.45)$$

where N_A and N_N are the numbers of apnoea and non-apnoea events respectively. For each i_x, the parameters k_x, r_x, sx_x and sy_x are tuned to minimise m as in (6.45). The parameters are optimised in a grid manner: each parameter is discretised over a range of possible values, and all combinations are tested. The interval between values was reduced until there was no significant difference in m between adjacent points on the grid.

6.3.3.4 Neural Network

A performance measure for S_{2n} is already defined, as the backpropagation algorithm optimises weights to minimise the MSE between actual and desired outputs, denoted E_{net} [Lippmann 1987]. However, given a system with a particular network architecture, there are many possible

networks with different weights, and the network with the smallest E_{net} does not always lead to the system with the best overall performance.

For apnoea detection, if the false negatives and false positives are zero then the E_{net} is also zero; however, given the inconsistency of the reference apnoeas, a zero error is almost impossible—distinguishing exactly between apnoea and non-apnoea events is very unlikely. The performance criterion of final output stage of the system is maximising SI(2) but, even though on average as E_{net} decreases SI(2) increases, the network with the minimum E_{net} almost never has the maximum (best performing) SI(2). Therefore, SI(2) is calculated at each iteration.

A common problem with neural networks is specialisation, where the network trains successfully and performs well with training data, but performs poorly when tested on other data [Lippmann 1987, Pao 1989, Pal and Mitra 1992]. This problem occurred with most initial trainings. To avoid specialisation, each training set is split into two by grouping every second event into one subset and the remainder into another subset. The network is trained by calculating E_{net} using one subset, and backpropagating the errors in the usual way. However, at each iteration of the backpropagation algorithm, the performance is tested on both sets by measuring SI(2). Because the events are taken from the same recordings, both subsets are similar in nature (they are not independent). Hence, if the network trains successfully but performs poorly on the whole set then it must have specialised. The system with the highest SI(2) is recorded as the result of the training, as opposed to the system with the lowest E_{net}.

6.3.4 Evaluation of System Performance and Experimental Results

This Section describes an example of training and testing a system. The system is tuned to maximise the performance criterion described in Section 6.3.3 relative to a set of training data. The various data sets are shown Table 6.3. The aim is to train the system using a reference standard that is as broad as possible, combining the opinion of experts, and hence the system is trained with B_1. The other data sets can then be used to test the system.

Set	Description	Apnoeas	Recordings
B_1		619	10
B_{1a}	agreed apnoeas	553	10
B_2	home recordings of normal infants, expert 4	28,339	85
B_3	home recordings of ALTE infants, experts 1 & 5	32,000	57

Table 6.3 Data sets which can be used for training and testing

One training of the system, using a training set of ten recordings or approximately 2000 events, typically takes from four to six weeks on an IBM Pentium 75 with 16MB of RAM, for the system implemented in the Modula2 language. As a result, only a small number of trainings are performed, and training with large data sets (> 4,000 events) is not computationally feasible.

The first stage of the system is S_{1f}. Using a detection system with a low false negative rate, five out of 619 apnoea are missed, and 2133 non-apnoea events are detected [Macey, et al. 1995].

This is a large reduction from approximately 150 hours of breathing signals, with less than 1% of apnoeas excluded. Thus, there is a total of 2847 events for which the properties measures are calculated.

Considering the stage S_{2m}, the parameters of the transformations for each p_x are optimised to minimise m, which is calculated for each i_x (see (6.45)). Compared to initial estimates, where the parameters were manually adjusted to maximise the separation of cumulative frequency distribution curves of i_{xA} and i_{xN} as viewed on screen, the optimised parameters result in a set of training inputs with which the overall system performance is the best of the systems tested.

Considering the neural network classifier S_{2n}, the most successful architectures have two hidden layers with either 7-5 or 8-5 nodes. Many architectures have been trained, but the two mentioned consistently outperform all others. As is standard with neural network training [Pao 1989], each architecture is trained multiple times (at least ten) with each set, starting with random initial weights. Once the network has ceased to improve for 10,000 iterations, restarting is used [Pao 1989], which occasionally allowed the network performance to improve. Each architecture often trains to a similar performance regardless of the initial weights; an exception is when the initial weights are wide ranging (between -5.0 and 5.0, as opposed to between approximately -1.0 and 1.0), and in these cases the network usually performs poorly.

Using SI(2) as a performance test each iteration means that the network that generalises most successfully in terms of then overall system performance is recorded. A pattern in all trainings is that initially both E_{net} and SI(2) improve, but, after 2,000 to 20,000 iterations, SI(2) stops improving while E_{net} is still decreasing. At this point the network is specialising, learning to classify only the training set events. Therefore, the network with the maximum SI(2) is selected as the resulting network; a training is considered finished once 10,000 iterations after restarting produce no further improvement in performance. Although calculating SI(2) each iteration more than doubles the time taken for each iteration (a typical training time is two days), it ensures that the final network has generalised.

The backpropagation training uses half the B_1 events; there is little difference between either half, but one consistently trained to slightly higher performances across all tested architectures. Some test trainings were performed with B_2, and a similar result is noticed: either the set of odd or the set of even numbered events consistently performed slightly better. Another variation tested is reducing the number of training events, and constructing a variety of training sets. However, the larger training sets lead to neural networks consistently generalising to a better overall performance.

The results are shown in Table 6.4. The system trained with B_1 is tested using B_2 and B_3. Testing on B_2, the results are not as good as for B_1 (Table 6.4). This is as expected since the system is trained on B_1, and the apnoeas in B_2 are based on a single expert's opinion. However, the system does detect a high percentage of apnoeas (94.1%) for only two false detections per three apnoeas detected, compared to the previous system with 12 false detections per 3 apnoeas detected. Performance with B_3 is similar to that with B_2. As mentioned previously, B_2 is not completely independent of the system as initial results were used as feedback to modify the system. Nevertheless, it is a large data set and if the system accurately detects 28,000 apnoeas then it is unlikely to have specialised. The B_3 results are similar to those with B_2. Note that B_1, B_2

and B_3 are different types of data sets—different experts performing the apnoea detection, different patient profiles, and different instruments. Thus, results between the different data sets cannot be directly compared. However, they can be used as indications of the extent to which the system has generalised.

System		False Negatives	False Positives
training data	test data	(% apnoeas detected)	(% of all events)
B_2	B_1	5.7 ± 0.01	41.1 ± 0.01
B_1	B_2	5.9 ± 0.01	39.1 ± 0.01
B_1	B_3	3.9 ± 0.01	40.8 ± 0.01
Previous System			
B_1	B_2	2 ± 1	80 ± 5

Table 6.4 Results for various systems system. The confidence intervals on the performance of the previous system are large because the figures were recorded and rounded by the experts, and the original results are no longer available.

As a test, independent training sets from B_2 are also used to train the network. While the B_2 training sets are larger, the resulting networks have almost the same performance as those trained using B_1 (see Table 6.4). Since the network performs equally well whether trained with B_1 or B_2, the training is generalising to independent data. This implies that there may be little advantage in attempting to improve the neural network training, for example by using faster training algorithms, different error calculations, or different combinations of training data.

In terms of clinical applications, the system is significantly improved over the previous one, which has been in daily use (Chapter 5). The false positives, while still significant, have been reduced (for the test data) from being 80% of the events detected to less 50%. Thus, for a night's recording with say 50 apnoeas, the previous detection method would detect approximately 250 events, whereas the current one would detect around 85 events. Given that clinicians look through every event, this represents a considerable improvement in system performance.

As a test, independent training sets from B_2 are also used to train the network. While the B_2 training sets are larger, the resulting networks have almost exactly the same performance as those trained using B_1 (see Table 6.4). Since the network performs equally well whether trained with B_1 or B_2, this confirms that the training is generalising to independent data. This implies that there may be little advantage in attempting to improve the neural network training, for example by using faster training algorithms, different error calculations, or different combinations of training data.

6.4 Discussion and Conclusions

In summary, a set of deterministic properties that discriminate between apnoea and similar non-apnoea events has been described, including mathematical descriptions giving measures of the properties. These measures are used as part of a new detection system, which is trained and tested using a variety of signals. Although it has a significant rate of false positives, the performance of

the new system is improved compared to previous systems. The performance is consistent across a range of test recordings.

The properties describe a small number of partially independent measures that discriminate between apnoea and non-apnoea events, and these properties are in line with the objectives outlined in Section 6.2.1. The property measures are more complex than the measures described by previous algorithms, and would not be suitable for analysing an entire breathing signal. The properties could be used effectively in conjunction with one of the apnoea detection algorithms that detects most apnoeas but that also has a high false positive rate. The key result is that the properties distinguish between apnoea and non-apnoea events in a manner that previous algorithms do not.

The tuning parameters allowed the properties to be optimised, maximising discrimination for a given training set. The window length for calculating standard deviation is a common parameter for three of the properties. Considering the smoothness, this property refers to the fine detail of the flat region, and the window L_4 is correspondingly short at 0.3 seconds. On the other hand, thinness is a characteristic of a wide region of a breathing signal, and is related to surrounding breathing. The optimum window duration for L_3 is 3.5 seconds, longer than the window lengths of flat regions and the flatness L_f and L_1, which are between L_3 and L_4 at 1.1 and 0.9 seconds respectively. Thus, the properties measure standard deviation with a range of window lengths, and this is one reason why they are relatively independent. The fact that window lengths optimised to values consistent with the property descriptions confirms the fact that the property measures are measuring what is described.

A consequence of the subjective interpretation is that there are very probably other properties that would meet the objectives in Section 6.2.1. Even using the same training data and the same experts, different descriptions could be developed. Given the same descriptions, different common features could be identified, and therefore different mathematical descriptions. The tuning parameters allow for some variation, but these are also subjectively added as what appeared to be the most useful parameters to adjust.

For the neural network, the backpropagation training combined with the standard feedforward architecture appears well-suited to the task of classification. It is noticeable that the performance of the system trained with B_1 is almost the same as the performance of the system trained with B_2: the false negative and false positive figures are within 1% of each other. The similarity of these results suggests that the technique of using half the reference events for training the neural network and the other half for testing is successful in avoiding specialisation. The consistent neural network performance also suggests that the backpropagation training method is adequate in terms of accuracy. There were no one-off exceptional results, nor any other indications of possible increased performance given the same i_x input values. The training times are the main area that the neural network could be improved, and there are many methods suggested in the literature. However, the training time of the neural network is less than the training time of other stages in the system.

The performance of the system is improved compared to previous algorithms, and it is consistent for different training and testing data (see Section 4.4) [Macey, et al. 1995]. A high percentage of apnoeas are detected (approximately 95%), but for a significant false positive rate

(approximately 40%). The system is objectively described, along with training criteria for each stage.

The 40% false positives means that the system does not have the same performance as human experts, and the main reason for this is the signal variation. Infants' physical breathing is variable, changing from variable amplitude, irregular length breaths to shallow amplitude, regular length breaths every 30 to 60 minutes. These differences in breathing relate to REM and quiet sleep stages that occur throughout a night's sleep. REM and quiet sleep can be observed in a Graseby breathing signal as regions of large and small ranges respectively [Tappin, et al. 1996a]. Another example of unusual events is when infants may almost have an apnoea. The baby stops breathing, takes a very quick breath, and then stops breathing a little longer, before resuming normal breathing. The infant has had an event in that their breathing pattern is interrupted, but the event is not a true apnoea. Overall, infants have a variety of behaviours, and classifying as either apnoea or non-apnoea is a simplification of the actual physical behaviour.

In terms of the goal of achieving the accuracy of human expert detection, this system is an improvement on previous work [Macey, et al. 1995]. The false positives, while still significant, have been reduced (for the test data) from being four fifths of the events detected to less than half. Expert input is still required to get a detection rate comparable to human experts, and so the system can be considered as an aid to a clinician, but not a replacement.

The breathing signal itself is also variable and not a direct measure of breathing. The abdomen and chest regions include artifact due to cardiac movement, the position of the sensor, the position of the infant (for example, below or on top of the sensor), and how securely the sensor is attached. Any movement of the infant causes noise, seen as dramatic large changes in amplitude. In addition, each instrument is calibrated differently: the Graseby needs recalibrating every three months, otherwise it can drift so far outside the normal range that it malfunctions [Tappin, et al. 1996a]. This is one reason why apnoea detection from a Graseby or similar signal has more inaccuracies than detection from an airflow or similar signal.

Future work could aim to improve the discrimination of properties. The more accurate the measures, the better the chance the system has of performing well. Another area is evaluating each apnoea in terms of importance. The high rate of false positives is mainly due to the performance goal of detecting *all* apnoeas. Some apnoeas are more significant than others, so if these could be distinguished, the system could be developed to focus on the important events. In any case, the limiting factor on the accuracy of the detection system is more likely to be the reference standard or the properties, and not the system design itself.

In conclusion, the expert system presented has been trained to detect apnoeas. The system and the training criteria are objectively described, and can be applied to a variety of signals. The systems that were trained have performances that are superior to a previous system. This is a first stage towards the ultimate goal of an apnoea detection algorithm that has similar accuracy to human experts.

Chapter 7

Conclusions

This chapter presents the overall conclusions of this thesis, as summarised from previous chapters. Some suggestions for future research are also presented.

Apnoea detection from an abdominal breathing signal is frequently performed, especially with data recorded from infants. However, mathematical descriptions of apnoea in terms of signal characteristics are lacking, as are accurate detection algorithms. Human expert opinion loosely defines what signal characteristics represent an apnoea, but expert opinion is inconsistent. Given the extent to which apnoea detection is used, there is a need for improved definitions and detection algorithms, and this thesis contributes to these areas.

There are some overall conclusions that can be drawn regarding apnoea detection from an abdominal or other similar signal. Firstly, apnoeas as represented within a breathing signal are not well defined. There are significant differences of opinion between experts even after discussing definitions and signal interpretations. Given that most human experts operate independently, there are likely to be high levels of disagreement when considering the opinions of a large number of experts. This has implications for comparing research, as results quoting apnoeas that have been detected in a particular study cannot necessarily be directly compared to results from other studies. The lack of a definitive agreement regarding a reference set of apnoeas means that mathematical definitions and detection algorithms cannot be accurate relative to all experts' opinions. They can only be evaluated relative to a reference that has inherent inconsistencies. These differences of opinion revolve around the details of the breathing signal. For example, a slight deviation on a flat region may be considered a breathing movement by one expert but not by another. There are other subtle differences in signal shape that cause differences in experts' interpretations.

When observing events that cause many of the disagreements, it is clear that many of these represent abnormal behaviour. Even though breathing may not have ceased completely, the signal does not represent normal breathing, and often appears to be a reduction in breathing. Given the physiological definition of an apnoea as a *cessation* of respiratory airflow, any slight breathing movement is sufficient for an event to not be classified as an apnoea. If an expert interprets a signal as having even a slight breathing movement, the event is rejected as an apnoea. As there is a range of signal deviations from large to almost imperceptible, an expert must at some point make a choice as to how large the deviation must be to represent a breathing movement, and this choice varies between experts.

The start and end points of an apnoea in a signal are also open to interpretation. Breathing is not a process that is either on or off, but instead airflow is continually increasing and decreasing. Thus, a breathing signal represents a range of airflows from almost nil to a full inhalation or exhalation. An expert decides exactly where in the signal breathing stopped or started, as there is no definitive beginning or end to a pause in breathing. The end is typically clearer than the start,

as there is no decay curve as at the start, but even so deciding the *exact* end requires the expert to interpret some point in the signal that represents a switch from no breathing to breathing.

There is a lack of documented apnoea detection algorithms, and thus there is a lack of systems with which to compare any new algorithm. Some previous detection algorithms were implemented and tested, and used as a reference against which to compare the new system.

The previous methods, whilst detecting the majority of apnoeas, have a high rates of false positives. The peak-to-peak and the standard deviation methods both appear to detect a great number of events that are flat, but that do not correspond to apnoeas. These include the events mentioned above that lead to disagreement between experts, but they also include many other events. Low amplitude breathing is a common problem, as the signal contains many short flat regions with each breath, and the amplitude of each breath is so low that the whole signal can be detected as flat, or not containing a peak. Poor signals, with erratic short flat periods and sharp peaks, are also detected as apnoeas. Both these events and low amplitude breathing signals are recognised by experts as non-apnoea events, but these simple detection algorithms do not distinguish them.

The peak-to-peak and standard deviation algorithms perform well in the task of discriminating between most of the breathing and apnoeas, and their computation times are short. Thus, they are useful for reducing the entire breathing recording to a number of events that include the majority of apnoeas. This is the basis of many existing apnoea detection algorithms, and they can be used as an aid to a clinician. With one large study involving over 400 nights of recorded breathing, the standard deviation algorithm was used to detect approximately 100,000 events. An expert viewed these events and accepted approximately 30,000 as apnoeas. Although this was still a time-consuming task, it would have been impractical for an expert to view the entire breathing records. These algorithms are therefore useful as preprocessors within a more sophisticated system for detection.

A mathematical description of apnoea has been developed that focuses specifically on discriminating between true apnoeas and problem events. Many previous methods of analysing breathing are designed to analyse an entire breathing signal and segment it into breaths. The events that cause problems for algorithm and expert detection are only a small component of the overall breathing signal, and as the methods are not specific to apnoea detection these events are not taken into account. A model of apnoea as a set of properties of a flat region forms a mathematical description of apnoea. Although the properties do not discriminate fully between all apnoea and non-apnoea events, they discriminate more than the previous algorithms using standard deviation and peak-to-peak. By designing the properties to discriminate specifically between apnoeas and the false positive events from the previous algorithms, the accuracy of the model is improved on what was previously available.

The cumulative frequency distribution curves of apnoea and non-apnoea events are different for each of the properties, a fact that can be observed and that is confirmed by Smirnov test values of almost zero. The properties are relatively independent, measuring different information.

An expert system for apnoea detection combines the new model of apnoea with the standard deviation detection algorithm to achieve improved detection performance. While there is still a significant rate of false positives, about 40%, this is greatly reduced from around 80% with

previous systems. The system does not achieve the accuracy of human experts, and in terms of clinical use, still requires an expert to verify events. However, the events presented to the expert are fewer in number: previously, with an 80% false positive rate, four out of five events detected would be false detections, compared with approximately two out of five events for the new system.

An important aspect of on-going research is the development of a reliable reference set of apnoea signals. A reference signal which represents an apnoea is defined by human experts, and therefore an improved measure of human expert opinion would lead to an improved representation of what is an apnoea. One possible approach would be to have experts rank events on a scale from unlikely but possible apnoea, to definite apnoea. The experts would not be constrained to decide between definite apnoea or definite non-apnoea, and would allow for uncertainty. There may still be a significant disagreement between experts in terms of the exact grading on the scale, but there would be likely to be fewer events that one expert would class as definitely apnoea and that another would class as definitely non-apnoea.

Definitions could be developed to discriminate between definite apnoea and definite non-apnoea events as ranked by the experts. The events about which there was uncertainty would not be considered as important to detect or reject. If the algorithm produced a measure of likelihood of an event being an apnoea, that measure could be compared against the experts' scaling. Overall, as such a reference would be a more accurate representation of human expert opinion, any definition or detection algorithm could be optimised to more closely represent the experts' opinion, which is the overall goal towards which this research is heading.

It is possible that a more accurate reference of apnoea for a abdominal breathing signal could be produced by experts viewing several breathing signals recorded at the same time. Thus, many of the events about which there could uncertainly if just viewing the one signal may be able to be easily classified by considering other signals.

The model presented is one solution out of a range of possibilities. It is likely that some other property measures would lead to improved discrimination, or improved independence between the properties. In terms of the characteristics of the four properties of a flat region (flatness, duration, thinness and smoothness), there may be other characteristics that lead to better discrimination. However, it is the mathematical descriptions of the characteristics that form the model, and so any change to the characteristics is only useful if it leads to new, improved mathematical measures.

The structure of the detection system presented appears suitable for the task of apnoea detection. The likely improvements are with components of the system. In particular, the properties are a key step as they give a measure that discriminates between events that previously were difficult to separate. The neural network classifier is well suited to this type of problem, where the exact form of the inputs may change with changing properties, and where the solution at this stage cannot meet the ultimate desired goal. It is possible that an improved classification system could be designed, or that the initial stage of detecting the flat regions could be more accurate, but these changes are not likely to affect the performance as significantly as the previous ones suggested.

In conclusion, this thesis highlights the need for a reference of human expert opinion of what signals represent apnoeas, signal definitions of apnoeas, and mathematically described algorithms. Towards these ends, this thesis describes a set of properties of apnoea signals that forms a model of apnoea, and as part of an expert system, that lead to improved accuracy of apnoea detection compared to previous algorithms. This work is a step towards accurate and reliable apnoea detection.

References

Abdulhamid, I., P. A. Vauthy, B. A. Barnett, D. R. Hufford, R. P. Reddy, and C. E. Hunt, "Comparison of 2-channel and 4-channel pneumograms," *Pediatric Pulmonology*, vol. **13**(4), pp. 245-9, 1992.

Abraham, N. G., V. A. Stebbens, M. P. Samuels, and D. P. Southall, "Investigation of cyanotic/apneic episodes and sleep-related upper airway obstruction by long-term non-invasive bedside recordings," *Pediatric Pulmonology*, vol. **8**, pp. 259-262, 1990.

Abreu e Silva, F. A., U. M. MacFadyen, A. Williams, and H. Simpson, "Sleep apnoea during upper respiratory infection and metabolic alkalosis in infancy," *Archives of Disease in Childhood*, vol. **61**, pp. 1056-1062, 1986a.

Abreu e Silva, F. A., A. Williams, and H. Simpson, "Sleep apnoea in infants with congenital stridor," *Archives of Disease in Childhood*, vol. **61**, pp. 1125-1137, 1986b.

Afonso, V. X., W. J. Thompkins, and J. G. Webster, "Quantitative measures of respiratory sinus arrhythmia for apnea detection," Proceedings of the 16th Annual international Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 94CH3474-4), IEEE, pp. 129-130, Baltimore, MD, USA, 3-6 November, 1994.

Ajmani, A., J. Mazumdar, and D. Jarvis, "Spectral analysis of an acoustic respiratory signal with a view to developing an apnoea monitor," *Australasian Physical and Engineering Sciences in Medicine*, vol. **19**(2), pp. 46-52, 1996.

Ariagno, R. L., C. Guilleminault, R. Korobkin, M. Owen-Boeddiker, and R. Baldwin, "'Near-miss' for sudden infant deah syndrome infants: a clinical problem," *Pediatrics*, vol. **71**(5), pp. 726-730, 1983.

Azimi-Sadjadi, M. R. and R.-J. Liou, "Fast learning process of multilayer neural networks using recursive least squares method," *IEEE Transactions on Signal Processing*, vol. **40**(2), pp. 446-450, 1992.

Barnard, E., "Optimization for training neural nets," *IEEE Transactions on Neural Networks*, vol. **3**(2), pp. 232-240, 1992.

Barnard, E. and D. Casasent, "A comparison between criterion functions for linear classifiers, with an application to neural nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. **19**(5), pp. 1030-1040, 1989.

Barnett, T. G., J. K. Stothers, and V. H. van Someren, "The detection and physiological evaluation of apnoeic episodes in human infants," *Journal of Physiology*, vol. **325**, p. 20P, 1981.

Barschdorff, D., A. Jaeger, and E. Trowitzsch, "Automatic assessment of an 'apnoea-severity-factor' combined with heart rate analysis during polysomnographic examinations in infants," Computers in Cardiology, IEEE, pp. 101-104, Bethesda, MD, USA, 25-28 September, 1994.

Baum, E. B. and D. Haussler, "What size net gives valid generalisation?," *Neural Computation*, vol. **1**, pp. 151-160, 1989.

Beauchamp, K. G., *Walsh Functions and their Applications*. London, New York: Academic Press, 1975.

Beckerman, R. C. and M. J. Wegmann, "A comparison of trachael breath sounds, airflow, and impedence pneumography in the detection of Childhood apnea," *Sleep*, vol. **8**(4), pp. 342-346, 1985.

Beckerman, R. C., M. J. Wegmann, and W. W. Waring, "Trachael breath sounds for detection of apnea in infants and children," *Critical Care Medicine*, vol. **10**(6), pp. 363-366, 1982.

Bello, M. G., "Enhanced training algorithms, and integrated training/architecture selection for multilayer perceptron networks," *IEEE Transactions on Neural Networks*, vol. **3**(6), pp. 864-874, 1992.

Beveridge, G. S. and R. S. Schechter, *Optimization: Theory and Practice*: McGraw Hill, 1970.

Biernacka, H. and N. J. Douglas, "Evaluation of a computerised polysomnography system," *Thorax*, vol. **48**, pp. 280-283, 1993.

Bliwise, D., N. G. Bliwise, H. C. Kraemer, and W. Dement, "Measurement error in visually scored electrophysiological data: respiration during sleep," *J Neurosci Methods*, vol. **12**(1), pp. 49-56, 1984.

BOC, *Operator's Manual Ohmeda Biox 3700 Pulse Oximeter*, BOC Health Care, Colorado, USA, Document 1118-301, 1986.

Bolton, D. P., B. J. Taylor, A. J. Campbell, B. C. Galland, and C. Cresswell, "Rebreathing expired gases from bedding: a cause of cot death?," *Archives of Disease in Childhood*, vol. **69**(2), pp. 187-190, 1993.

Bolton, D. P. G., E. A. S. Nelson, B. J. Taylor, and I. L. Weatherall, "Thermal balance and prone sleep position: results of theoretical modeling," Conference program of the Fourth SIDS International Conference, SIDS International, p. 177, Washington, USA, 23-26 June, 1996.

Brent, R. P., "Fast training algorithms for multilayer neural nets," *IEEE Transactions on Neural Networks*, vol. **2**(3), pp. 346-353, 1991.

Brigham, E. O., *The FFT and its Applications*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

Brooks, J. G., "Apnea of infancy and sudden infant death syndrome," *Am J Dis Child*, vol. **136**, pp. 1012-1023, 1982.

Brooks, J. G., "Apparent life-threatening events and apnea of infancy," *Clinics in Perinatology*, vol. **19**(4), pp. 809-838, 1992.

Brouillette, R., D. Hanson, R. David, L. Klemka, A. Szatkowski, S. Fernbach, and C. Hunt, "A diagnostic approach to suspected obstructive sleep apnea in children," *Journal of Pediatrics*, vol. **105**(July), pp. 10-14, 1984.

Brouillette, R. T., S. K. Fernbach, and C. E. Hunt, "Obstructive sleep apnea in infants and children," *Journal of Pediatrics*, vol. **100**(1), pp. 31-40, 1982.

Brouillette, R. T., A. S. Morrow, D. E. Weese-Mayer, and C. E. Hunt, "Comparison of respiratory inductive plethysmography and thoracic impedance monitoring," *Journal of Pediatrics*, vol. **111**, pp. 377-383, 1987.

Brown, P. J., R. Dove, B. Price, S. Fong, and R. P. K. Ford, "Continuous multiple location body temperature measurement on infants," IEEE Engineering in Medicine and Biology Annual Conference, pp. 1052-1053, 1990.

Brown, P. J., R. A. Dove, C. S. Tuffnell, and R. P. K. Ford, "Oscillations of body temperature at night," *Archives of Disease in Childhood*, vol. **67**, pp. 1255-1258, 1992.

Bruckert, R., F. Perrin, J. Pernier, and M. J. Challamel, "[Methode de monitoring respiratoire et de detection automatique des apnees chez le nourrisson a risque de mort subite][French]," *Medical and Biological Engineering and Computing*, vol. **20**(Nov.), pp. 693-698, 1982.

Burgess, R. C., "Computerized polysomnographic analysis systems," *Journal of Clinical Neurophysiology*, vol. 7(1), pp. 145-154, 1990.

Burr, D. J., "Experiments on neural net recognition of spoken and written text," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. **36**(7), pp. 1162-1168, 1988.

Butcher-Puech, M. C., D. J. Henderson-Smart, D. Holley, J. L. Lacey, and D. A. Edwards, "Relation between apnoea duration and type and neurological status of preterm infants," *Archives of Disease in Childhood*, vol. **60**, pp. 953-958, 1985.

Chakrabarti, S., N. Bindal, and K. Theagharajam, "Robust radar target classifier using artificial neural networks," *IEEE Transactions on Neural Networks*, vol. 6(3), pp. 760-766, 1995.

Cornwell, A. C. and S. Laxminarayan, "Sleep apnea in "near miss" and control infants," *IEEE ninth annual conference of the engineering in medicine and biology society*, pp. 1111-1112, 1987.

Corometrics, *Operator's Manual Model 500 Infant Monitor*, Corometrics Medical Systems, Connecticut, USA, Manual P/N 1135AA-01, 1985.

Corwin, M. J., N. Neuman, J. Silvestri, D. Crowell, S. Davidson-Ward, L. Brooks, C. Hunt, G. Lister, M. Willinger, and CHIME, "Accuracy of apnea scoring with the CHIME monitor: comparison to polysomnography," Conference Program of The Fourth SIDS International Conference59, Washington, USA, June 23-26, 1996.

David, H. A., "Order-Statistics," *Wiley*, 1981.

de Villiers, J. and E. Barnard, "Backpropagation neural nets with one and two hidden layers," *IEEE Transactions on Neural Networks*, vol. **4**(1), pp. 136-141, 1992.

Delivoria-Papadopoulos, M., N. P. Poncevic, and F. A. Oski, "Postnatal changes in oxygen transport of term, preterm, and sick infants: the role of red cell 2,3-diphosphoglycerate and adult haemoglobin," *Pediatric Research*, vol. **5**, pp. 235-245, 1971.

Department of Health, *Recording Child Health and Development*. Wellington, New Zealand: Department of Health, 1984.

Douglas, N. J., S. Thomas, and M. Jan, "Clinical value of polysomnography," *Lancet*, vol. **339**(Feb.), pp. 347-350, 1982.

Dove, R., J. Brown, R. Fright, C. Tuffnell, and R. Ford, "Computer polygraphic system for infants at risk for sudden infant death syndrome (SIDS)," *Australasian Physical and Engineering Sciences in Medicine*, vol. **13**, pp. 188-191, 1990.

Dove, R. A., (.1988.), *Instrumentation for Paediatric Cardio-respiratory Assessment*, Electrical and Electronic Engineering, University of Canterbury, Christchurch, New Zealand

Dunne, K. P., M. McKay, and T. G. Mathews, "'Near miss' sudden infant death and obstructive apnoea," *Archives of Disease in Childhood*, vol. **61**, pp. 1039-1040, 1986.

East, K. A. and T. D. East, "Computerized acoustic detection of obstructive apnea," *Computer Methods and Programs in Biomedicine*, vol. **21**, pp. 213-220, 1985.

Felten, E. W., O. Martin, S. W. Otto, and J. Hutchinson, "Multi-scale training of a large backpropagation net," *Biological Cybernetics*, vol. **62**, pp. 503-509, 1990.

Fletcher, R. and C. M. Reeves, "Function minimization by conjugate gradients," *Computer Journal*, vol. 7, pp. 149-154, 1964.

Fogel, D. B., "An information criterion for optimal neural network selection," *IEEE Transactions on Neural Networks*, vol. **2**(5), pp. 490-497, 1991.

Ford, R. P. K., "Postneonatal mortality in Christchurch," *The New Zealand medical journal*, vol. **99**(815), pp. 939-941, 1986.

Ford, R. P. K., P. J. Brown, R. A. Dove, C. S. Tuffnell, and P. M. Macey, "HomeLog: long term recording of infant temperature, respiratory and cardiac signals in the home environment," *Journal of Paediatrics and Child Health*, Suppl. 1, pp. 26-33, 1992.

Ford, R. P. K., J. Larkin, S. Hart, and D. M. Tappin, "Infant home apnoea monitors in Christchurch: an audit," *New Zealand Medical Journal*, vol. **107**(Jan), pp. 12-13, 1994.

Ford, R. P. K., C. S. Tuffnell, P. M. Macey, T. M. Tappin, and M. Sambamoorthy, "Rectal temperature changes and apnea," Conference Program of the Fourth SIDS International Conference, SIDS International, p. 123, Washington, USA, June 23-26, 1996.

Franks, C. I., D. M. Johnston, and B. H. Brown, "Non-invasive home monitoring of respiratory patterns in infants," *Develop. Med. Child Neurol.*, vol. **19**, pp. 748-756, 1977.

Ghitza, O., "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. **2**(1), pp. 115-132, 1994.

Ghorbani, A. A. and N. C. Bhavsar, "Incremental communication - reduced accuracy for multilayer neural networks," *IEEE Transactions on Neural Networks*, vol. **6**(6), pp. 1375-1385, 1995.

Gibson, E., "Apnea," in *Intensive Care of the Fetus and Neonate*, A. R. Spitzer Ed. St. Louis: Mosby-Year Book, 1996a, p. 470.

Gibson, E. R., "Sudden infant death syndrome," in *Intensive Care of the Foetus and Neonate*, A. R. Spitzer Ed. St. Louis, USA: Mosby-Year Book, Inc., 1996b, pp. 482-493.

Giles, C. L. and T. Maxwell, "Learning, invariance, and generalization in high-order neural networks," *Applied Optics*, vol. **26**(23), pp. 4972-4978, 1987.

Gordon, D., R. J. Cohen, D. Kelly, S. Akselrod, and D. C. Shannon, "Sudden Infant Death Syndrome: abnormalities in short term fluctuations in heart rate and respiratory activity," *Pediatric Research*, vol. **18**(10), pp. 921-926, 1984.

Gordon, D., D. P. Southall, D. H. Kelly, A. Wilson, S. Akselrod, J. Richards, B. Kenet, R. Kenet, R. J. Cohen, and D. C. Shannon, "Analysis of heart rate and respiratory patterns in sudden infant death syndrome victims and control infants," *Pediatric Research*, vol. **20**, pp. 680-684, 1986.

Gotman, J., "Automatic seizure detection: improvements and evaluation," *Electroencephalography and clinical Neurophysiology*, vol. **76**, pp. 317-324, 1990.

Gotman, J., P. Gloor, and N. Schaul, "Comparison of traditional reading of the EEG and automatic recognition of interictal epileptic activity," *Electroencephalography and Clinical Neurophysiology*, vol. **44**, pp. 44-60, 1978.

Graseby Medical Ltd., (.1988.), "Respiration Monitor Type MR10 Technical Service Manual No. SM 108," . Colonial Way, Watford, Herts., WD2 4LG, England, p. 17.

Griggs, S. D., D. M. Tappin, R. P. K. Ford, and M. P. Wailoo, "Overnight oscillations of rectal temperature," *Archives of Disease in Childhood*, vol. **73**, p. 277, 1995.

Guilleminault, C., R. Ariagno, R. Korobkin, S. Coons, M. Owen-Boeddiker, and R. Baldwin, "Sleep parameters and respiratory variables in 'near miss' sudden infant death syndrome infants," *Pediatrics*, vol. **68**, pp. 354-360, 1981.

Guilleminault, C., J. v. d. Hoed, and M. M. Mitler, "Clinical overview of the sleep apnea syndromes," in *Sleep apnea syndromes*, C. Guilleminault and W. C. Dement Eds. New York: Alan R. Liss, 1978, pp. 1-12.

Guilleminault, C., R. Peraita, M. Souquet, and W. C. Dement, "Apneas during sleep in infants: Possible relationship with Sudden Infant Death Syndrome," *Science*, vol. **190**, pp. 677-699, 1975.

Guilleminault, C., M. Souquet, R. L. Ariagno, R. Korobkin, and F. B. Simmons, "Five cases of near-miss sudden infant death syndrome and development of obstructive sleep apnea syndrome," *Pediatrics*, vol. **73**(1), pp. 71-78, 1984.

Guo, H. and S. B. Gelfand, "Analysis of gradient descent learning algorithms for multilayer feedforward neural networks," *IEEE Transactions on Circuits and Systems*, vol. **38**(8), pp. 883-894, 1991.

Gyulay, S., D. Gould, B. Sawyer, D. Pond, A. Mant, and N. Saunders, "Evaluation of a microprocessor-based portable home monitoring system to measure breathing during sleep," *Sleep*, vol. **10**(2), pp. 130-142, 1987.

Haidmayer, R., T. Kenner, and R. Kurz, "Paradoxical ventilatory response of babies to pure oxygen (author's trans.) [Ger]," *Wiener Medizinische Wocherschrift*, vol. **130**(3), pp. 128-129, 1980.

Haidmayer, R., R. Kurz, T. Kenner, H. Wurm, and K. P. Pfeiffer, "Physiological and clinical aspects of respiration control in infants with relation to the sudden infant death syndrome," *Klinische Wochenschrift*, vol. **60**, pp. 9-18, 1982.

Harper, R. M., B. Leake, T. Hoppenbrouwers, M. B. Sterman, D. J. McGinty, and J. E. Hodgman, "Polygraphic studies of normal infants and infants at risk for the Sudden Infant Death Syndrome: heart rate and variability as a function of state," *Paediatric Research*, vol. **12**, pp. 778-785, 1978.

Harper, R. M., V. L. Schechtman, and K. A. Kluge, "Machine classification of infant sleep state using cardiorespiratory measures," *Electroencephalography and Clinical Neurophysiology*, vol. **67**, pp. 379-387, 1987.

Haykin, S., (.1996.), "Neural networks expand SP's horizons," in *IEEE Signal Processing Magazine*, pp. 24-49.

Henderson-Smart, D. J. and G. Cohen, "Apnoea in the newborn infant," *Australasian Paediatric Journal*, vol. **Suppl.**, pp. 63-66, 1986.

Hewertson, J., C. F. Poets, M. P. Samuels, S. G. Boyd, B. G. R. Neville, and D. P. Southall, "Epileptic seizure-induced hypoxemia in infants with apparent life-threatening events," *Pediatrics*, vol. **94**, pp. 148-156, 1994.

Hill, J. R. and K. A. Rahimtulla, "Heat balance and the metabolic rate of new-born babies in relation to environmental temperature; and the effect of age and weight on metabolic rate," *Journal of Physiology*, vol. **180**, pp. 239-265, 1965.

Hodgman, J. E., T. Hoppenbrouwers, S. Geidel, A. Hadeed, M. B. Sterman, R. Harper, and D. McGinty, "Respiratory behaviour in near-miss sudden infant death syndrome," *Pediatrics*, vol. **69**, pp. 785-792, 1982.

Hoppenbrouwers, T., R. M. Harper, J. E. Hodgman, M. B. Sterman, and D. J. McGinty, "Polygraphic studies of normal infants during the first six months of life. II. Respiratory rate and variability as a function of state," *Paediatric Research*, vol. **12**(2), pp. 120-125, 1978.

Hoppenbrouwers, T., J. E. Hodgman, R. M. Harper, E. Hofmann, M. B. Sterman, and D. J. McGinty, "Polygraphic studies of normal infants during the first six months of life: III. Incidence of apnea and periodic breathing," *Pediatrics*, vol. **60**(4), pp. 418-425, 1977.

Hoppenbrouwers, T., J. E. Hodgman, R. M. Harper, and M. B. Sterman, "Respiration during the first six months of life in normal infants: IV. Gender differences," *Early Human Development*, vol. **4**(2), pp. 167-177, 1980a.

Hoppenbrouwers, T., J. E. Hodgman, D. McGinty, R. M. Harper, and M. B. Sterman, "Sudden Infant Death Syndrome: Sleep apnea and respiration in subsequent siblings," *Pediatrics*, vol. **66**(2), pp. 205-214, 1980b.

Hunt, C. E. and R. T. Brouillette, "Sudden infant death syndrome: 1987 perspective," *Journal of Pediatrics*, vol. **110**(5), pp. 669-677, 1987.

Hunt, C. E., R. T. Brouillette, and D. Hanson, "Apnea-onset definition significantly affects pneumogram results," *Sleep*, vol. **11**(3), pp. 286-90, 1988.

Hunt, C. E., R. T. Brouillette, D. Hanson, R. J. David, I. M. Stein, and M. Weissbluth, "Home pneumograms in normal infants," *Journal of Pediatrics*, vol. **106**(April), pp. 551-555, 1985a.

Hunt, C. E., R. T. Brouillette, K. Liu, and L. Klemka, "Day-to-day pneumogram variability," *Pediatric Research*, vol. **19**, pp. 174-177, 1985b.

Iiguni, Y. and H. Sakai, "A real-time learning algorithm for a multilayered neural network based on the extended Kalman filter," *IEEE Transactions on Signal Processing*, vol. **40**(4), pp. 959-966, 1992.

Iiguni, Y., H. Sakai, and H. Tokumaru, "A real-time learning algorithm for a multilayered neural network based on the extended Kalman filter," *IEEE Transactions on Signal Processing*, vol. **40**(4), pp. 959-966, 1992.

Jeffrey, H., R. A. Cunningham, A. Cubis, and D. J. C. Read, "New methods to separate artifacts from normal and defective breathing patterns in different sleep-states, if infants are

monitored at home," *Australian and New Zealand Journal of Medicine*, vol. **11**, pp. 406-411, 1981.

Kahn, A., D. Blum, E. Rebuffat, M. Sottiaux, J. Levitt, A. Bochner, M. Alexander, J. Grosswasser, and M. F. Muller, "Polysomnographic studies of infants who subsequently died of sudden infant death syndrome," *Pediatrics*, vol. **82**, pp. 721-727, 1988.

Kahn, A., J. Groswasser, E. Rebuffat, M. Sottiaux, D. Blum, M. Foerster, P. Franco, A. Bochner, M. Alexander, A. Bachy, P. Richard, M. Verghote, D. L. Polain, and L. Wayenberg, "Sleep and cardiorespiratory characteristics of infant victimes of sudden death: a prospective case-control study," *Sleep*, vol. **15**(4), pp. 287-292, 1992.

Katz-Salomon, M. and J. Milerad, "Ventilatory and heart-rate responses to moderate CO_2-loading in infants at risk of SIDS," Conference Program of the Fourth SIDS International Conference, SIDS International, pp. 128-129, Washington, USA, 23-26 June, 1996.

Kelly, D. H., H. Golub, D. Carley, and D. C. Shannon, "Pneumograms in infants who subsequently died of sudden infant death syndrome," *Journal of Pediatrics*, vol. **109**, pp. 249-254, 1986.

Kelly, D. H., A. M. Walker, L. Cahen, and D. C. Shannon, "Periodic breathing in siblings of sudden infant death syndrome victims," *Pediatrics*, vol. **66**(4), pp. 515-520, 1980.

Kempe, C. H., H. K. Silver, and D. O'Brien, *Current paediatric diagnosis and treatment*, 3rd ed. Los Altos, California: Lange Medical Publications, 1974.

Kendrick, A. H., N. Wiltshire, and J. R. Catterall, "Scoring of apnoeas during sleep," *Thorax Proceedings British Thoracic Society*, 1990.

Kim, H. and K. Nam, "Object recognition of one-DOF tools by a back-propagation neural network," *IEEE Transactions on Neural Networks*, vol. **6**(3), pp. 484-487, 1995.

Kirjavainen, T., D. Cooper, O. Polo, and C. E. Sullivan, "The static-charge-sensitive bed in the monitoring of respiration during sleep in infants and young children," *Acta Paediatrica*, vol. **85**, pp. 1146-52, 1996.

Kollias, S. and D. Anastassiou, "An adaptive least squares algorithm for the efficient training of artificial neural networks," *IEEE Transactions on Circuits and Systems*, vol. **36**(8), pp. 1092-1101, 1989.

Kuan, C.-M. and K. Hornik, "Convergence of learning algorithms with constant learning rates," *IEEE Transactions on Neural Networks*, vol. **2**(5), pp. 484-489, 1991.

Laxminarayan, S., O. Mills, L. Michelson, A. C. Cornwell, A. Marmarou, E. F. Costigan, Jr., and E. D. Weitzman, "Sudden infant death syndrome: a digital computer-based apnoea monitor," *Medical and Biological Engineering and Computing*, vol. **21**(March), pp. 191-196, 1983.

Laxminarayan, S., O. Mills, L. Michelson, A. C. Cornwell, A. Marmarou, and E. D. Weitzman, "A digital computer application in the on-line monitoring of expired CO_2 in sudden infant death syndrome studies," Applications of Computers in Medicine (IEEE Cat. No. TH0095-0), IEEE, pp. 117-127, 1982.

Lee, D., R. Caces, K. Kwiatkowski, D. Cates, and H. Rigatto, "A developmental study on types and frequency distribution of short apnoeas (3 to 15 seconds) in term and preterm infants," *Pediatric Research*, vol. **22**(3), pp. 344-349, 1987.

Leverich, M. K., J. L. Silberberg, and S. Weininger, "Design and development of a multi-channel signal-acquisition system for recording an apnea waveform database," Seventh Annual IEEE Symposium on Computer-Based Medical Systems (Cat. No.94CH2426-4), IEEE, pp. 213-216, Winston-Salem, USA, 10-12 June, 1994.

Lindgren, B. W., "11.2.3 Comparisons of Distributions," in *Statistical Theory*, third ed. New York: MacMillan Publishing Co., Inc., 1976, pp. 494-496.

Lippmann, R. P., (.1987.), "An introduction to computing with neural nets," in *IEEE ASSP Magazine*, p. 4.

Lippmann, R. P., (.1989.), "Pattern classification using neural networks," in *IEEE Communications Magazine*, pp. 47-64.

Macey, P. M., R. P. K. Ford, P. J. Brown, J. Larkin, R. W. Fright, and K. Garden, "Apnoea detection: human performance and reliability of a computer algorithm," *Acta Paediatrica*, vol. **84**, pp. 1103-1107, 1995.

Macey, P. M., R. P. K. Ford, and J. S. J. Li, "Reliable apnea detection from an abdominal breathing signal," Presented at: the Fourth SIDS International Conference, Washington, USA, June 23-26, 1996a.

Macey, P. M., J. S. J. Li, and R. P. K. Ford, "Designing an expert system for apnoea detection," Proceedings of the Third New Zealand Conference of Postgraduate Students in Engineering and Technology83-88, University of Canterbury, Christchurch, July 1-2, 1996b.

Macey, P. M., J. S. J. Li, and R. P. K. Ford, "Deterministic properties of apnoea in an abdominal breathing signal," *Medical and Biological Engineering and Computing*, under revision, 1998.

Macey, P. M., J. S. J. Li, and R. P. K. Ford, "Expert system for apnoea detection," *Engineering Applications of Artificial Intelligence*, accepted January 1998.

MacFadyen, U. M., G. Borthwick, H. Simpson, M. McKay, and J. Neilson, "Monitoring for central apnoea in infancy - limitations of single channel recordings," *Archives of Disease in Childhood*, vol. **63**, pp. 282-287, 1988.

Markel, *IEEE Transactions on Audio and Electroaccoustics*, vol. **AU-20**(Dec.), pp. 168, 1972.

Marshall, R. J., "The determination of peaks in biological waveforms," *Computers and Biomedical Research*, vol. **19**(4), pp. 319-329, 1986.

Martin, R. J., M. J. Miller, and W. A. Carlo, "Pathogenesis of apnea in preterm infants," *Journal of Pediatrics*, vol. **109**(5), pp. 733-741, 1986.

Mason, J. R., R. M. Harper, and R. F. Pacheo, "Analysis of respiratory data during sleep and waking," Proceedings of the Digital Equipment Users Society, pp. 567-571, San Diego, California, November, 1974.

Mayotte, M. J., J. G. Webster, and W. J. Tompkins, "Reduction of motion artifacts during paediatric / infant apnoea monitoring," *Medical and Biological Engineering and Computing*, vol. **34**(Jan), pp. 93-96, 1996.

McClave, J. T. and P. G. Benson, *Statistics for Business and Economics*, 5 ed. Singapore: Maxwell MacMillan International, 1991.

Mehrotra, K. G., C. K. Mohan, and S. Ranka, "Bounds on the number of samples needed for neural learning," *IEEE Transactions on Neural Networks*, vol. **2**(6), pp. 548-558, 1991.

Michalopoulou, Z. H., L. W. Nolte, and D. Alexandrou, "Performance evaluation of multilayer perceptrons in signal detection and classification," *IEEE Transactions on Neural Networks*, vol. **6**(2), pp. 381-386, 1995.

Miles, L. E., Clinical Monitoring Center, and Vitalog Monitoring, "Use of the Vitalog "lunch-box" home monitor for evaluation of obstructive sleep apnoea," in *Sleep and Health Risk*, J. H. Peter, T. Penzel, T. Podszus, and P. van Wichert Eds. New York: Springer-Verlag, 1989, 10 pages.

Milner, A. D. and N. Ruggins, "Sudden infant death syndrome: recent focus on the respiratory system," *British Medical Journal*, vol. **298**(18 March), pp. 689-690, 1989.

Mitchell, E. A., J. M. Brunt, and C. Everard, "Reduction in mortality from sudden infant death syndrome in New Zealand: 1986-1992," *Archives of Disease in Childhood*, vol. **70**, pp. 291-294, 1994.

Mitchell, E. A., R. Scragg, A. W. Stewart, D. M. O. Becroft, B. J. Taylor, R. P. K. Ford, I. B. Hassall, B. M. J. Barry, E. M. Allen, and A. P. Roberts, "Results from the first year of the New Zealand cot death study," *The New Zealand Medical Journal*, vol. **104**(906), pp. 71-76, 1991.

Mitchell, E. A., B. J. Taylor, R. P. K. Ford, A. W. Stewart, D. M. O. Becroft, J. M. D. Thompson, R. Scragg, I. B. Hassall, D. M. J. Barry, E. M. Allen, and A. P. Roberts, "Four modifiable and other major risk factors for cot death: The New Zealand study," *Journal of Paediatric and Child Health*, vol. **28** Suppl. 1, pp. S3-8, 1992.

Mitchell, I., R. P. C. Barclay, R. Railton, and e. al, "Frequency and severity of apnoea in lower respiratory tract infection in infancy," *Archives of Disease in Childhood*, vol. **58**, pp. 497-499, 1983.

Mood, A. M., F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, 3 ed. New York: McGraw-Hill, 1974.

Moody, J. O. and P. J. Antsaklis, "The dependence identification neural network construction algorithm," *IEEE Transactions on Neural Networks*, vol. **7**(1), pp. 3-15, 1996.

Moyles, T. P., R. F. Erlandson, and T. Roth, "A nonparametric statistical approach to breath segmentation," IEEE Engineering in Medicine and Biology Society 11th Annual Conference, IEEE, pp. 330-331, Seattle, USA, 9-12 November, 1989.

Mussell, M. J., "the need for standards in recording and analysing respiratory sounds," *Medical and Biological Engineering and Computing*, vol. **30**(March), pp. 129-139, 1992.

National Institutes of Health, "Infantile Apnea and Home Monitoring," US Dept. Health & Human Sciences Public Health Service, National Institutes of Health, 1987.

Nekovei, R. and Y. Sun, "Back propagation network and its configuration for blood vessel detection in angiograms," *IEEE Transactions on Neural Networks*, vol. **6**(1), pp. 64-72, 1995.

Nelson, E. A. S., S. M. Williams, B. J. Taylor, B. Morris, and R. P. K. Ford, "Postneonatal mortality in south New Zealand: necropsy data review," *Paediatric and Perinatal Epidemiology*, vol. **3**, pp. 375-385, 1989.

Neter, J., W. Wasserman, and G. A. Whitmore, *Applied Statistics*, 1978.

Nilsson, N. J., *Learning machines: foundations of trainable pattern classifying systems*, 1965.

Ning, T. and J. D. Bronzino, "Automatic classification of respiratory signals," Images of the Twenty-First Century. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, pp. 669-670, Seattle, USA, 9-12 November, 1989.

Noll, M. A., "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. **41**(Feb.), pp. 293-309, 1967.

Oppenheim, A. V., R. H. Schafer, and T. G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proceedings of the IEEE*, vol. **56**(8), pp. 1264-1291, 1968.

Oren, J., D. Kelly, and D. C. Shannon, "Identification of a high-risk group for Sudden Infant Death Syndrome among infants who were resuscitated for sleep apnea," *Pediatrics*, vol. 77(4), pp. 495-499, 1986.

Pal, S. K. and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," *IEEE Transactions on Neural Networks*, vol. **3**(5), pp. 683-697, 1992.

Pao, Y. H., *Adaptive Pattern Recognition and Neural Networks*, 1989.

Park, D. J., B. E. Jun, and J. H. Kim, "Novel fast training algorithm for multilayer feedforward neural network," *Electronics Letters*, vol. **28**(6), pp. 543-544, 1992.

Parmalee, A. H., E. Stern, and M. A. Harris, "Maturation of respiration in prematures and young infants," *Neuropadiatrie*, vol. **3**, pp. 294-304, 1972.

Peirano, P., B. B. Singh, J. Lacombe, H. Cherrier, and N. Monod, "Cardiorespiratory patterns during sleep and wakefulness in infants who subsequently died of SIDS," IEEE Engineering in Medicine and Biology Society 10th Annual Conference, IEEE, pp. 1930-1931, New Orleans, USA, 4-7 November, 1988.

Pfeiffer, K. P., R. Haidmayer, P. Kerschhaggl, R. Kurz, and T. Kenner, "Statistical evaluation of the respiratory pattern as a risk factor for the sudden infant death syndrome," *Meth. Inform. Med.*, vol. **23**(1), pp. 41-46, 1984.

Poets, C. F., V. A. Stebbens, M. P. Samuels, and D. P. Southall, "The relationship between bradycardia, apnea, and hypoxemia in preterm Infants," *Pediatric Research*, vol. **34**(2), pp. 144-147, 1993.

Ponsonby, A.-L., T. Dwyer, L. E. Gibbons, J. A. Cochrane, M. E. Jones, and M. J. McCall, "Thermal environment and sudden infant death syndrome: case-control study," *British Medical Journal*, vol. **304**(Feb.), pp. 277-282, 1992.

Powell, M. J. D., "Restart procedures for the conjugate gradient method," *Mathematical Programming*, vol. **12**, pp. 241-254, 1977.

Qin, S.-Z., H.-T. Su, and T. J. McAvoy, "Comparison of four neural net learning methods for dynamic system identification," *IEEE Transactions on Neural Networks*, vol. **3**(1), pp. 122-130, 1992.

Rakowski, S., A. Smith, K. Prowse, and M. Allen, "Computer based monitoring and analysis of sleep apnoea," Progress Reports on Electronics in Medicine and Biology, Institute of Electronic and Radio Engineers, pp. 1-8, 1986.

Rawson, D., S. A. Peterson, and M. P. Wailoo, "Rectal temperature of normal babies after first diphtheria, pertussis, and tetanus immunisation`," *Archives of Disease in Childhood*, vol. **65**, pp. 1305-1307, 1990.

Revow, M. D., S. J. England, and H. O'Beirne, "Robust computer algorithm for detecting breaths in noisy ventilatory waveforms from infants," *Medical and Biological Engineering and Computing*, vol. **24**(Nov.), pp. 609-615, 1986.

Reyneri, L. M. and E. Filippi, "Modified backpropagation algorithm for fast learning in neural networks," *Electronics Letters*, vol. **26**(19), pp. 1564-1566, 1990.

Richards, J. M., J. R. Alexander, E. A. Shinebourne, M. de Swiet, A. J. Wilson, and D. P. Southall, "Sequential 22-hour profiles of breathing patterns and heart rate in 110 full-term infants during their first 6 months of life," *Pediatrics*, vol. **74**, pp. 763-777, 1984.

Ruck, D. W., S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, vol. **1**(4), pp. 296-298, 1990.

Rumelhart, D. E. and J. L. McClelland, *Parallel distributed processing*, vol. 1-2. Cambridge, MA: MIT Press, 1986.

Sahakian, A. V. and W. J. Tompkins, "A multi-microcomputer-based neonatal apnea monitor," 10th Annual Northeast Bioengineering Conference, IEEE, pp. 151-156, Hanover USA, 1982.

Scalero, R. S. and N. Tepedelenliogu, "A fast new algorithm for training feedforward neural networks," *IEEE Transactions on Signal Processing*, vol. **40**(1), pp. 202-210, 1992.

Schafer and Rabiner, *Journal of the Acoustical Society of America*, vol. **47**(2), pp. 265, 1970.

Schechtman, V. L., R. M. Harper, K. A. Kluge, A. J. Wilson, H. J. Hoffmann, and D. P. Southall, "Cardiac and respiratory patterns in normal infants and victims of the Sudden Infant Death Syndrome," *Sleep*, vol. **11**(5), pp. 413-424, 1988.

Schechtman, V. L., R. M. Harper, K. A. Kluge, A. J. Wilson, and D. P. Southall, "Correlations between cardiorespiratory measures in normal infants and victims of Sudden infant Death Syndrome," *Sleep*, vol. **13**(4), pp. 304-317, 1990.

Schechtman, V. L., R. M. Harper, A. J. Wilson, and D. P. Southall, "Sleep apnea in infants who succumb to the Sudden Infant Death Syndrome," *Pediatrics*, vol. **87**(6), pp. 841-846, 1991.

Scheffer, F., H. Stute, D. Sontheimer, A. Meissner, and O. Linderkamp, "Are numbers of apneas, bradycardias or tachycardias indicators for a higher risk of SIDS in preterm infants with BPD?," Conference Program of the Fourth SIDS International Conference, SIDS International, pp. 130-131, Washington, USA, 23-26 June, 1996.

Schluter, B., D. Buschatz, and E. Trowitzsch, "Apnea characteristics of children who later died: comparison of sudden infant death with other causes of death [German]," *Wien Med Wochenschr*, vol. **146**, pp. 13-14, 321-323, 1996.

Scholten, C. A. and J. E. Vos, "Descriptors of the rhythmicity in respiration and heart beat of newborn infants," *Medical and Biological Engineering and Computing*, vol. **19**, pp. 83-90, 1981.

Scholten, C. A., J. E. Vos, and H. F. R. Prechtl, "Compiled profile of respiration, heart beat and motility in newborn infants: a methodological approach," *Medical and Biological Engineering and Computing*, vol. **23**(1), pp. 15-22, 1985.

Schroeter, J. and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. **2**(1), pp. 133-150, 1994.

Siegwart, D. K., L. Tarassenko, S. J. Roberts, J. R. Stradling, and J. Partlett, "Sleep apnoea analysis from neural network post-processing," Artificial Neural Networks, IEE Conference Publication No. 409, pp. 427-432, 26-28 June, 1995.

Skadberg, B. T. and T. Markestad, "Differences in behaviour, CO_2 rebreathing, and physiological responses after getting the head covered by bedding during prone and supine sleep," Conference Program of the Fourth SIDS International Conference, SIDS International, p. 178, Washington, USA, 23-26 June, 1996.

Sokolov, R. T. and J. C. Rogers, "Time-domain cepstral transformations," *IEEE Transactions on Signal Processing*, vol. **41**(3), pp. 1161-1169, 1993.

Sorsa, T., H. Koivo, and H. Koivisto, "Neural networks in process fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. **21**(4), pp. 815-825, 1991.

Southall, D. P., "Role of apnoea in the sudden infant death syndrome: a personal view," *Pediatrics*, vol. **80**, pp. 73-84, 1988.

Southall, D. P., J. M. Richards, K. C. Lau, and E. A. Shinebourne, "An explanation for failure of impedance apnoea alarm systems," *Archives of Disease in Childhood*, vol. **55**, pp. 63-65, 1980.

Southall, D. P., J. M. Richards, V. Stebbens, A. J. Wilson, V. Taylor, and J. R. Alexander, "Cardiorespiratory function in 16 full-term infants with sudden infant death syndrome," *Pediatrics*, vol. **78**(5), pp. 787-796, 1986.

Sprott, J., *The Cot Death Cover-up?*: Penguin Books (NZ) Ltd., 1996.

Stein, I. M. and D. C. Shannon, "The pediatric pneumogram: a new method for detecting and quantitating apnea in infants," *Pediatrics*, vol. **55**(5), pp. 599-603, 1975.

Stein, I. M., A. White, J. L. Kennedy, R. L. Merisalo, H. Chernoff, and J. B. Gould, "Apnea recordings of healthy infants at 40, 44, and 52 weeks postconception," *Pediatrics*, vol. **63**(5), pp. 724-730, 1979.

Storck, K., M. Karlsson, and P. Ask, "Heat transfer evaluation of the nasal thermistor technique," *IEEE Transactions on Biomedical Engineering*, vol. **43**(12), pp. 1187-1191, 1996.

Sturman, S., *Apnoea Detection Using a Neural Network*, University of Canterbury, Christchurch, November 1991.

Takagi, H., N. Suzuki, T. Koda, and Y. Kojima, "Neural networks designed on approximate reasoning architecture and their applications," *IEEE Transactions on Neural Networks*, vol. **3**(5), pp. 752-760, 1992.

Talbot, S. A. and U. Gessner, *Systems Physiology*. New York: Wiley, 1973.

Tappin, D. M., R. P. Ford, K. P. Nelson, B. Price, P. M. Macey, R. Dove, J. Larkin, and B. Slade, "Breathing, sleep state, and rectal temperature oscillations," *Archives of Disease in Childhood*, vol. **74**, pp. 427-431, 1996a.

Tappin, D. M., R. P. K. Ford, K. Nelson, B. Price, P. Macey, and R. Dove, "Central apnoea is not increased in normal infants after vaccination," Conference Program of The Fourth SIDS International Conference, SIDS International & SIDS Alliance, p. 119, Washington, USA, June 23-26, 1996b.

Tappin, D. M., R. P. K. Ford, K. P. Nelson, B. Price, P. M. Macey, and R. Dove, "The febrile stress of routine vaccination does not increase central apnoea in normal infants," *Acta Paediatrica*, vol. **86**, pp. 873-880, 1997.

Thach, B. T., "Sleep apnea in infancy and childhood," *Med Clin North America*, vol. **69**, pp. 1289-1315, 1985.

Tonkin, S. and T. Gunn, "The role of upper airway structure and function in SIDS," Conference Program for the Fourth SIDS International Conference, SIDS International, p. 171, Washington, USA, 23-26 June, 1996.

Tonkin, S. L., J. H. Stewart, and S. Withey, "Obstruction of upper airway as a mechanism for sudden infant death: evidence for a restricted nasal airway contributing to pharyngeal obstruction," *Sleep*, vol. **3**(4), pp. 375-382, 1980.

Tribolet, J. M., "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. **ASSP-25**(2), pp. 170-177, 1977.

Tudehope, D. I., Y. M. Rogers, Y. R. Burns, H. Mohay, and m. J. O'Callaghan, "Apnoea in very low birthweight infants: outcome at 2 years," *Australasian Paediatric Journal*, vol. **22**, pp. 131-134, 1986.

Tuffnell, C. S., (.1993.), *Biomedical Engineering Aspects of Infant Thermoregulation and Respiration*, Electrical and Electronic Engineering, University of Canterbury, Christchurch, 160 pages.

Upton, C. J., A. D. Milner, and G. M. Stokes, "Combined impedance and inductance for the detection of apnoea of prematurity," *Early Human Development*, vol. **24**, pp. 55-63, 1990.

Valdes-Dapena, M. A., "Sudden Infant Death Syndrome: a review of the medical literature 1974 - 1979," *Pediatrics*, vol. **66**(4), pp. 597-614, 1980.

Vandenplas, Y., *Oesophageal pH monitoring for gastro-oesophageal reflux in infants and children*. Chichester, England: Wiley, 1992.

Vogl, T. P., J. K. Mangis, A. K. Rigler, W. T. Zink, and D. L. Alkon, "Accelerating the convergence of the back-propagation method," *Biological Cybernetics*, vol. **59**, pp. 257-263, 1988.

Wailoo, M., S. Peterson, H. Whittaker, and P. Goodenough, "Sleeping body temperature in 3-4 month old infants," *Archives of Disease in Chldhood*, vol. **64**, pp. 596-599, 1989.

Ward, S. L. D., T. G. Keens, L. S. Chan, B. E. Chipps, S. H. Carson, D. D. Deming, V. Krishna, H. M. MacDonald, G. I. Martin, K. S. Meredith, T. A. Merritt, B. G. Nickerson, R. A. Stoddard, and A. L. v. d. Hal, "Sudden infant death syndrome in

infants evaluated by apnea programs in California," *Pediatrics*, vol. **77**(4), pp. 451-458, 1986.

Werthammer, J., J. Krasner, J. DiBenedetto, and A. R. Startk, "Apnea monitoring by acoustic detection of airflow," *Pediatrics*, vol. **71**(1), pp. 53-55, 1983.

West, J. B., R. M. Peters Jnr., G. Aksnes, K. L. Maret, J. S. Milledge, and R. B. Schone, "Nocturnal periodic breathing at altitudes of 6,300 and 8,050 m," *Journal of Applied Physiology*, vol. **61**, pp. 280-287, 1986.

Whyte, K. F., M. B. Allen, M. F. Fitzpatrick, and N. J. Douglas, "Accuracy and significance of scoring hypopneas," *Sleep*, vol. **15**(3), pp. 257-60, 1992.

Widrow, B. and M. A. Lehr, "30 years of adaptive neural networks: perceptron, madaline, and backpropagation," *Proceedings of the IEEE*, vol. **78**(9), pp. 1415-1441, 1990.

Widrow, B., R. G. Winter, and R. A. Baxter, "Layered neural nets for pattern recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. **36**(7), pp. 1109-1118, 1988.

Wilkie, R. A., M. H. Bryan, S. Gaston, and A. C. Bryan, "Maturation of peripheral chemoreceptors in normal newborn infants," Federation Proceedings, p. 823, 1987.

Wilks, P. A. D. and M. J. English, "Accurate segmentation of respiration waveforms from infants enabling identification and classification of irregular breathing patterns," *Medical Engineering and Physics*, vol. **16**, pp. 19-23, 1994.

Wilks, P. A. D. and M. J. English, "A system for rapid identification of respiratory abnormalities using a neural network," *Medical Engineering and Physics*, vol. **17**(7), pp. 551-555, 1995.

Wilson, A. J., C. I. Franks, and I. L. Freeston, "Algorithms for the detection of breaths from respiratory waveform recordings of infants," *Medical and Biological Engineering and Computing*, vol. **20**(May), pp. 286-292, 1982.

Wilson, A. J., V. Stevens, C. I. Franks, and D. P. Southall, "Analysis of long-term cardiorespiratory recordings from infants who subsequently suffered SIDS," *Annals New York Academy Sciences*, vol. **533**(8), pp. 390-410, 1988.

Wintrobe, M. M., G. W. Thorn, R. D. Adams, E. Braunwald, K. J. Isselbacher, and R. G. Petersdorf, *Harrison's Principles of Internal Medicine*, Seventh ed. Tokyo: McGraw-Hill Kogakusha, 1974.

Yount, J. E., "Technical problems in recognising and monitoring infant apnea," Proceedings of the IEEE Engineering in Medicine and Biology Society 11th Annual Conference pp. 325-326, 1989.

Yu, X.-H. and S.-X. Cheng, "Training algorithms for backpropagation neural networks with optimal descent factor," *Electronics Letters*, vol. **26**(20), pp. 1698-1700, 1990.

Zoldac, J. T., O. Soykan, M. R. Neuman, and C. S. Group, "An electronic simulator for testing infant apnea monitors that utilizes realistic physiologic data," Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, pp. 661-662, San Diego, 1993.